Cumhuriyet International Journal of Education – Cumhuriyet Uluslararası Eğitim Dergisi
e-ISSN: 2147-1606
Volume / Cilt 9 | Issue / Sayı 2 June / Haziran 2020

Monte Carlo Simülasyon Yönteminde Tekrar Sayısı Klasik Test Kuramı Parametreleri İçin Kaç Olmalıdır?

Duygu Koçak[1]

Type/Tür:
Research/Araştırma
Received/Geliş Tarihi: August 29/ 29 Ağustos 2019
Accepted/Kabul Tarihi: February 20/ 20 Şubat 2020
Page numbers/Sayfa No: 410-429
Corresponding Author/İletişimden Sorumlu Yazar:
duygu.kocak@alanya.edu.tr

Öz

Son yıllarda yapay veri ile yapılan çalışmaların sayısı giderek artmaktadır. Yapay veriler birçok modelin, istatistiksel tekniğin, kuramın test edilmesinde dolayısıyla geliştirilmesinden önemli role sahiptir. Simülasyon çalışmalarındaki tekrar sayısının gerçeği yansıtan sonuçlar üretmedeki önemi tartışılmazdır. Monte Carlo simülasyon yöntemi kullanılarak bir araştırma tasarlandığında, tekrar sayısı araştırma sonuçlarının güvenilirliği ve geçerliliği için çok önemlidir. Ancak, simülasyonda kaç tekrarın yeterli olduğu konusunda net bir bilgi yoktur. Bu çalışmada, Klasik Test Kuramı temelinde Monte Carlo simülasyon yöntemindeki tekrar sayısının madde ve test parametresi tahminlerine etkisini belirlemek ve gerekli tekrar sayısını belirlemek amaçlanmıştır. Bu amaçla, farklı koşullar altında tekrar sayısının değiştirilmesiyle elde edilen veriler toplam varyans oranı, Cronbach Alfa katsayısı, madde madde ortalama ortalaması ve model veri uyumu parametreleri incelenmiştir. Bu çalışma bir Monte Carlo simülasyon çalışmasıdır. Araştırmada veri üretimi ve analizi için R programı (2011) "psych" paketi kullanılmıştır. Bu çalışmada, tek boyutlu bir yapıdaki madde sayısı 20'ye, cevap kategorisi 5'e sabitlenerek, örneklem büyüklüğü 100, 250, 500, 1000 ve 3000 olarak değiştirilmiştir. Çalışmanın sonuçlarına göre, Klasik Test Kuramı'na dayalı bir çalışmada araştırmacıların, benzer koşullarda örneklem büyüklüğü 100 iken 1000 tekrar ile, örneklem büyüklüğü 250 iken 500 tekrar ile, örneklem 500 iken 250 tekrar ile ve örneklem büyüklüğü 1000 ve 3000 iken 100 tekrar ile veri üretmeleri önerilmektedir.

Anahtar Kelimeler: Klasik Test Kuramı, simülasyon, Monte Carlo, tekrar sayısı.

Suggested APA Citation /Önerilen APA Atıf Biçimi:
Koçak, D.(2020). Monte Carlo simülasyon yönteminde tekrar sayısı klasik test kuramı parametreleri için kaç olmalıdır? *Cumhuriyet International Journal of Education*, *9*(2), 410-429. http://dx.doi.org/10.30703/cije.613114

[1] Dr. Öğr. Üyesi, Alanya Alaaddin Keykubat Üniversitesi, Eğitim Bilimleri Bölümü, Antalya/Türkiye
Assist. Prof. Dr. Alanya Alaaddin Keykubat University, Department of Educatioanl Sciences, Antalya/ Turkey
e-mail: duygu.kocak@alanya.edu.tr **ORCID ID:** orcid.org/0000-0003-3211-0426

What should be the number of replications in Monte Carlo Simulation Method for Classical Test Theory Parameters?

Abstract

The importance of the number of repetitions in the simulation studies to produce truth-reflecting results is indisputable. When a research is designed using Monte Carlo simulation technique, the number of repetitions is very important for the reliability and validity of the research results. However, there is no clear information on how many repetitions are sufficient. In this study, it is aimed to determine the effect of number of repetitions in Monte Carlo simulation method on item and test parameter estimations in Classical Test Theory and to determine the number of repetitions required. For this purpose, the data obtained by changing the number of replication under different conditions total variance ratio Cronbach's Alpha coefficient average of item discrimination and model-data-fit parameters were examined. This study is a Monte Carlo simulation study. In the research, R program "psyc" package was used for data generation and analysis. In this study, the number of items in a one-dimensional structure is fixed to 20, the response category is 5, and the sample size is changed to 100, 250, 500, 1000 and 3000. According to results of the study, in a study based on CTT, it is suggested that researchers produce data with 1000 replications when sample size is 100, 500 replications when sample size is 250, 250 replications when sample size is 500 and 100 replications when sample size is 1000 and 3000.

Keywords: Classical Test Theory, simulation, Monte Carlo, replication number.

Giriş

Eğitim ve psikoloji alanında bilişsel ve duyuşsal olan soyut özellikler doğrudan gözlenemeyeceği için çoğunlukla bir test ya da ölçek aracılığıyla ölçülmektedir. Bireyin testi oluşturan maddelere verdiği yanıtlar yani tepkiler aracılığıyla ilgili özelliğe dair çıkarımda bulunulmaktadır. Bu yolla bir özelliğin ölçülmesi amacıyla geliştirilen test ve bu testin uygulanması sonucu elde edilecek veriler anlamlandırılırken bir test kuramı temel alınmalıdır. Bu noktada Klasik Test Kuramı (KTK) varsayımlarının kolay karşılanır olması nedeniyle en sık tercih edilen test kuramıdır (Çelen, 2008; De Ayala, 2009).

Klasik Test Kuramı en yaygın kullanılan test kuramı olmasına karşın çeşitli sınırlılıkları bulunmaktadır. Bu sınırlılıklar beraberinde, bu sınırlılıkları ortadan kaldıracak ya da minimize edecek koşulların neler olabileceği sorularını getirmektedir. Örneğin bir ölçeğin belirli bir gruba uygulanması sonucu elde edilen veri setinde, rastgele kayıp mekanizmasına sahip yüksek oranda kayıp veri bulunması durumunda hangi kayıp veri baş etme yöntemi kullanılırsa geçerlik ve güvenirlik en az etkilenir sorusuyla karşılaşılabilmektedir. Benzer olarak, yeni geliştirilen istatistiksel bir tekniğin hangi koşullar altında nasıl performans gösterdiğinin araştırılması, güçlü ve zayıf yönlerinin bilinmesini sağlayacaktır ancak araştırmalarda ele alınan değişkenler ve bu değişkenlerin planlanan koşulları, düzeyleri her zaman gerçek veriler ile elde edilememektedir. Bu neden araştırmalar çoğu zaman yapay veriler kullanılarak yani simülasyon çalışmaları ile yapılmaktadır. Simülasyon yöntemi, dağılım özellikleri ve parametreleri belli olan veriler üreterek, çeşitli istatistiksel yöntemlerin denenebilmesi ve yöntemlerin performansının karşılaştırılabilmesine olanak sunar (Hauck ve Anderson, 1984). Simülasyon, istatistiksel tekniklerin belirli koşullardaki performanslarını belirlemek, farklı yöntemleri karşılaştırmak amacıyla, özellikleri belli veri setleri üreterek bunlar üzerinde denemeler yapmaktır (Sobol, 1971).

Simülasyon çalışmaları ele alınan ve canlandırılan durumlara bağlı olarak farklı isimler almaktadır. Zamana bağlı olarak değişiklik gösteren durum ve koşulların simülasyonu dinamik, zamana bağlı olarak değişiklik göstermeyen durumların simülasyonu statik, matematiksel bir model ile tanımlanabilen olayların simülasyonu deterministik, bir model ile tam olarak tanımlanamayan olayların simülasyonu ise stokastik simülasyon olarak adlandırılmaktadır ve bunlar arasında en sık kullanılan simülasyon türü stokastik simülasyondur. Stokastik simülasyonda Monte Carlo yöntemi ile rastgele sayılar üretilir (Naylor, Blantify, Burdick ve Chu, 1968), bu yolla önceden belirlenmiş özelliklere sahip veriler üretilerek çeşitli istatistiksel analiz yöntemlerinin belirlenen koşullardaki performansı incelenir. Stokastik simülasyon ve Monte Carlo yöntemleri bu yönüyle teknik bir deneme çalışmasıdır (Sobol, 1971) ve bir durumun simülatif verilerle incelenmesine simülasyon, model örneklemesi ya da Monte Carlo adı verilmektedir (Rubinstein, 1981). Harwell, Stone, Hsu ve Kirisci (1996), bir test kuramı ile ilgili Monte Carlo çalışması yapılırken izlenmesi gerekenleri ise sekiz adımda özetlemiştir:

- Çalışmanın amacını yansıtan araştırma sorusu ya da soruları belirlenir
- Değişkenler ve koşulları (düzeyleri) tanımlanır
- Uygun deneysel tasarım oluşturulur
- Belirlenen koşullar bir test kuramı temel alınarak üretilir
- Parametreler kestirilir
- Karşılaştırma yapılır
- Bu işlem tasarımdaki her hücre için tekrarlanır
- Elde edilen sonuçlar hem çıkarımsal hem betimsel olarak değerlendirilir.

Bu sonuçlar araştırma sorularına yanıt oluşturur aynı zamanda.

Bu basamaklar simülasyon çalışmasında izlenen basamakları eğitim araştırmaları açısından ifade etmektedir. Son yıllarda eğitim araştırmalarında simülasyon çalışmalarının yaygınlaşması nedeniyle simülasyon çalışması basamaklarına, prosedürlerine ilişkin çalışmalara da önem artmaktadır. Leventhall ve Ames (2019), National Council on Measurement in Education (NCME) 2019 yılı buluşmasında yaptıkları Madde Tepki Kuramı'nda Monte Carlo simülasyon çalışmaları için SAR kullanımı isimli çalıştaylarında, simülasyon çalışmalarının eğitim araştırmalarındaki önemine dikkat çekmişlerdir. Aynı organizasyonda Madde Tepki Kuramı temelli test simülasyonları için yazılım paketleri: WinGen3, SimulCAT, MSTGen, and IRTEQ (Yoo, Han ve Oh, 2019) isimli bir diğer çalıştayda da Bireye Uyarlanmış Test (BUT), test parametrelerinin kestirimleri, test eşitleme gibi konularda simülatif veri üretmek için kullanılabilecek yazılımları ve uygulamaları ele alınmıştır. Ölçme ve değerlendirme alanındaki önemli organizasyonlardan biri olan NCME'de simülasyon ve Monte Carlo yöntemine dikkat çekilmektedir. Bu durumun simülasyon çalışmalarının giderek önem kazanmasının sonucu olduğu düşünülebilir. NCME 2018 yılı buluşmasında toplam 91 çalışmada, 2019 yılı buluşmasında ise toplam 77 çalışmada simülasyon yapıldığı belirtilmiştir. Bu çalışmalarda, test eşitleme, yapısal eşitlik modeli, Genellenebilirlik Kuramı, Klasik Test Kuramı, Madde Tepki Kuramı sıklıkla ele alınıp, çeşitli koşullarda madde ve test parametreleri kestirilerek karşılaştırılmıştır.

Sobol (1971), Monte Carlo yönteminde, tanımlanan modelde belirlenen koşulları yansıtacak veri seti elde etmek için her biri diğerinden bağımsız olacak

şekilde N defa sürecin tekrarlanacağını ve böylece N farklı sayıda örneklem üretileceğini ifade etmektedir. Buna göre Monte Carlo yönteminde tekrar sayısı artırılarak varyans minimum hale getirilmektedir ve sonuçlara en az hata ile ulaşılmaktadır. Veri üretme aşamasında yetersiz sayıda tekrar, tahmin kestirimlerde hataya yol açmaktadır (Brooks, 2002; Gifford ve Swaminathan, 1990; Stone, 1993; Hammersly ve Handscombe, 1964; Hutchinson ve Bandalos, 1997; Lewis ve Orav, 1989). Buna göre Monte Carlo çalışmasında tekrar sayısını artırmak kestirimlerde hatayı azaltacaktır ancak tekrar sayısının kaç olması gerektiğine dair net bir açıklama yapılamamaktadır (Hutchinson ve Bandalos, 1997). Mundform, Schaffer, Kim, Shaw ve Thongteeraparp (2011), simülasyon çalışmalarında tekrar sayısını belirlemede bir prosedürün olmadığını bu nedenle tekrar sayısının araştırmacının insiyatifinde olduğunu ifade etmektedir. Binois, Huang, Gramary ve Ludkovsk (2019), Monte Carlo yönteminde, örneklem büyüdükçe gerekli tekrar sayısının daha az olduğunu ve bunun tüm simülasyon yöntemlerin ortak noktası olduğunu belirtmektedirler. Buna göre küçük örneklemlerde daha çok tekrara, büyük örneklemlerde ise daha az tekrara ihtiyaç duyulduğu ifade edilebilir (Harwell, Rubinstein, Hayes ve Olds, 1992). Alanyazında 10000 ile 10 arasında değişen tekrar sayıları ile yapılan çalışmalar hatta hiç tekrar olmadan yapılan çalışmalar bulunmaktadır (Fay ve Gerow; 2013; ; Glen Satten, Flanders ve Yang, 2001; Hambleton, Jones ve Rogers, 1993; Harwell ve Janosky, 1991; Hulin, Lissak ve Drasgow, 1982; Kannan, Sgammato, Tannenbaum ve Katz, 2015; Kéry ve Royle, 2016; Murie ve Nadon, 2018; Saeki ve Tango, 2014; Qualls ve Ansley, 1985; Yen, 1987).

Simulasyon çalışmalarının gerçeği yansıtır sonuçlar üretmesinde tekrar sayısının önemi tartışmasızdır ve Monte Carlo yöntemi kullanılarak bir araştırma tasarlanması durumunda tekrar sayısı araştırma sonuçlarının güvenirliği ve geçerliği açısından oldukça önemlidir fakat kaç tekrarın yeterli olacağına ilişkin net bir bilgi bulunmamaktadır. Bu çalışmada Monte Carlo yönteminde tekrar sayısının, Klasik Test Kuramı'nda madde ve test parametresi kestirimlerine etkisinin ve gerekli tekrar sayısının belirlenmesi amaçlanmıştır. Bu amaçla farklı örneklem büyüklüklerinde tekrar sayısı değişimlenerek hangi koşulda en az kaç tekrara ihtiyaç duyulduğu, açıklanan toplam varyans oranı, Cronbach Alfa katsayısı, madde ayırt edicilikleri ortalaması ve model veri uyumu parametreleri kestirilerek belirlenmiştir.

Yöntem

Bu araştırmada Klasik Test Kuramı temel alınarak Monte Carlo yöntemi ile üretilen verilerde tekrar sayısının test ve madde parametrelerine etkisi ve gerekli tekrar sayısının belirlenmesi amaçlanmıştır. Bu haliyle olsaydı ne olurdu sorusuna cevap arayan bir Monte Carlo simülasyon çalışmasıdır (Dooley, 2002).

Verilerin Üretilmesi ve Analizi

Araştırmada Monte Carlo yönteminde tekrar sayısının Klasik Test Kuramı madde ve test parametrelerine etkisini belirlemek amacıyla, veriler, örneklem büyüklüğü 100, 250, 500, 1000 ve 3000, madde sayısı 20, yanıt kategorisi çoklu (1,2,3,4 ve 5) olacak şekilde R programı "psych" paketi kullanılarak üretilmiştir ve analiz edilmiştir. Araştırmada etkisi araştırılan temel değişken olan tekrar sayısı ise 5, 10, 25, 50, 100, 250, 500 ve 1000 ve 10000 olarak değişimlenmiştir. Veriler, tek boyutlu ve açıklanan toplam varyans oranı %40 olarak, Cronbach Alfa iç tutarlılık anlamında güvenirlik

katsayısının ise 0,70 olarak belirlenmiştir. Buna göre Monte Carlo yöntemiyle tekrar sayısı değişimlenerek üretilen veriler Açımlayı Faktör Analizi (AFA) ile analiz edilmiş ve elde edilen açıklanan toplam varyans oranı kestirilmiş ve %40 kriter alınarak yorumlanmıştır. Aynı zamanda Cronbach Alfa katsayısı hesaplanmış ve elde edilen değerler veri üretme aşamasında belirlenen 0,70 değeri baz alınarak yorumlanmıştır. Üretilen verilerde madde parametrelerine ilişkin herhangi bir manipülasyon yapılmamıştır. Her bir veride madde ayırt edicilikleri ve buna bağlı olarak testin ayırt edicilik ortalaması hesaplanarak betimsel olarak yorumlanmıştır.

Üretilen verilere Doğrulayıcı Faktör Analizi (DFA) yapılarak $\chi 2$, AIC ve RMSEA değerleri kestirilmiştir. Veriler tek boyutlu olacak şekilde manipüle edilerek üretildiğini için DFA yapılırken maddelerin tümü tek bir faktöre tanımlanarak analiz gerçekleştirilmiş ve verinin modele uyumu kestirilmiştir. Aynı örneklem büyüklüğünde farklı tekrar sayıları kullanılarak üretilen verilere DFA yapılarak ele edilen $\chi 2$ uyum istatistiklerinin karşılaştırılmasında $\chi 2$ değerleri arasındaki farkın manidarlığından faydalanılmıştır. AIC istatistiğinin değerlendirilmesinde daha küçük değer elde edilen modelin veriye daya uyumlu olduğu değerlendirilmesi yapılmaktadır. Bu nedenle aynı örneklem büyüklüğünde farklı tekrarlarla elde edilen verilerden kestirilen AIC değerlerinin karşılaştırılmasında ise betimsel karşılaştırma yoluna gidilmiştir. RMSEA değerinin yorumlanmasında ise Tabachnick ve Fidell (2007) tarafından belirtilen aralıklara göre değerlendirme yapılmıştır. Buna göre $0 <$ iyi uyum $\leq 0,05$ ve $0,05 <$ kabul edilebilir uyum $\leq 0,08$ aralıklarına göre değerlendirme yapılmıştır.

Bulgular

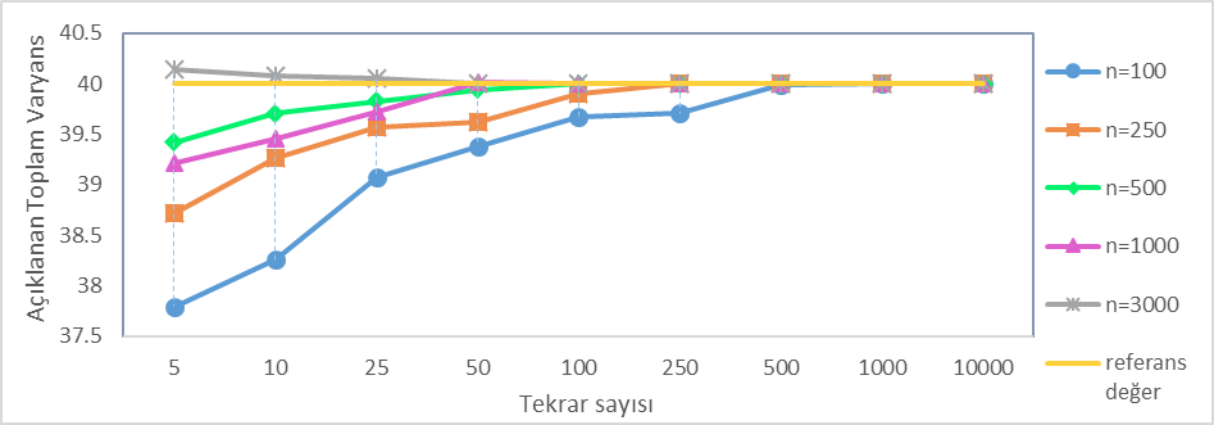
Monte Carlo yönteminde tekrar sayısının Klasik Test Kuramı'nda açıklanan toplam varyans oranına etkisinin belirlenmesi amacıyla farklı tekrar sayılarında veriler üretilerek AFA yapılmış ve açıklanan toplam varyans oranı kestirilmiştir. Sonuçlar Tablo 1'de yer almaktadır.

Tablo 1
Açıklanan Toplam Varyans Oranının Tekrar Sayısına Bağlı Değişimi

Tekrar sayısı	**Açıklanan toplam varyans oranı (%)**				
	N=100	N=250	N=500	N=1000	N=3000
5	37.79	38.72	39.42	39.21	40.14
10	38.26	39.26	39.71	39.45	40.08
25	39.07	39.57	39.83	39.72	40.05
50	39.38	39.62	39.94	40.01	40.00
100	39.67	39.97	40.00	40.00	40.00
250	39.71	40.00	40.00	40.00	40.00
500	39.99	40.00	40.00	40.00	40.00
1000	40.01	40.00	40.00	40.00	40.00
10000	40.00	40.00	40.00	40.00	40.00

Tablo 1 incelendiğinde örneklem büyüklüğü arttıkça hedeflenen referans değere (bu çalışma için 40%) ulaşmak için daha az tekrara ihtiyaç duyulduğu

görülmektedir. Şekil 1 her bir örneklem büyüklüğünde tekrar sayısına bağlı olarak elde değişen açıklanan toplam varyans oranlarını sunmaktadır



Şekil 1. Farklı örneklem büyüklüklerinde açıklanan toplam varyans oranının (%) tekrar sayısına bağlı değişimi

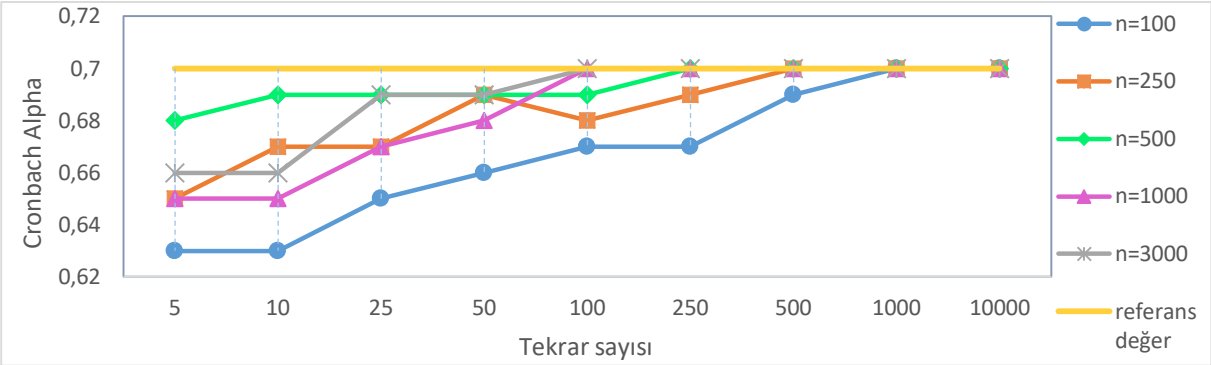
Şekil 1 incelendiğinde referans değer olan %40 açıklanan varyans oranına, örneklem büyüklüğü 100 olduğunda 1000 tekrar sayısında erişildiği, örneklem büyüklüğü 250 olduğunda 250 tekrar sayısıyla erişildiği, örneklem büyüklüğü 500 ve 1000 olduğunda 100 tekrar ile, örneklem büyüklüğü 3000 olduğunda ise 50 tekrar ile erişildiği görülmektedir. Bu tekrar sayılarından sonra elde edilen kestirimlerin belirlenen referans değerde sabitlendiği görülmektedir. Örneklem büyüklüğü arttıkça veri üretme aşamasında belirlenen açıklanan toplam varyans oranına daha az sayıda tekrar ile erişildiği görülmektedir.

Tablo 2 her bir örneklem büyüklüğü için farklı tekrar sayılarında kestirilen Cronbach Alfa katsayısını içermektedir. Veri üretme aşamasında Cronbach Alfa değeri 0.70 olarak sınırlandırılmıştır bu nedenle üretilen verilerden elde edilen değerin de 0.70 olması beklenmektedir. Bu değere ulaşılan tekrar sayısı o örneklem büyüklüğü için yeterli tekrar sayısı olarak belirlenmiştir.

Tablo 2
Cronbach Alfa Katsayısının Tekrar Sayısına Bağlı Değişimi

Tekrar sayısı	**Cronbach Alfa**				
	N=100	N=250	N=500	N=1000	N=3000
5	0,63	0,65	0,68	0,65	0,66
10	0,63	0,67	0,69	0,65	0,66
25	0,65	0,67	0,69	0,67	0,69
50	0,66	0,69	0,69	0,68	0,69
100	0,67	0,68	0,69	0,70	0,70
250	0,67	0,69	0,70	0,70	0,70
500	0,69	0,70	0,70	0,70	0,70
1000	0,70	0,70	0,70	0,70	0,70
10000	0,70	0,70	0,70	0,70	0,70

Tablo 2 incelendiğinde örneklem büyüklüğü arttıkça belirlenen Cronbach Alfa değerine (0.70) daha az tekrar sayısı ile erişildiği görülmektedir. Her bir örneklem büyüklüğünde tekrar sayısına bağlı olarak Cronbach Alfa katsayısındaki değişim Şekil 2'de sunulmuştur.



Şekil 2. Farklı örneklem büyüklüklerinde Cronbach Alfa katsayısının tekrar sayısına bağlı değişimi

Şekil 2 ve Tablo 2 incelendiğinde örneklem büyüklüğü 100 olduğunda 1000 tekrar, 250 olduğunda 500 tekrar, 500 olduğunda 250 tekrar, 1000 ve 3000 olduğunda ise 100 tekrar ile belirlenen Cronbach Alfa değerine (0.70) erişildiği görülmektedir. Buna göre örneklem büyüklüğü arttıkça veri üretmen aşamasında belirlenen değere erişmek için gerekli tekrar sayısı azalmaktadır.

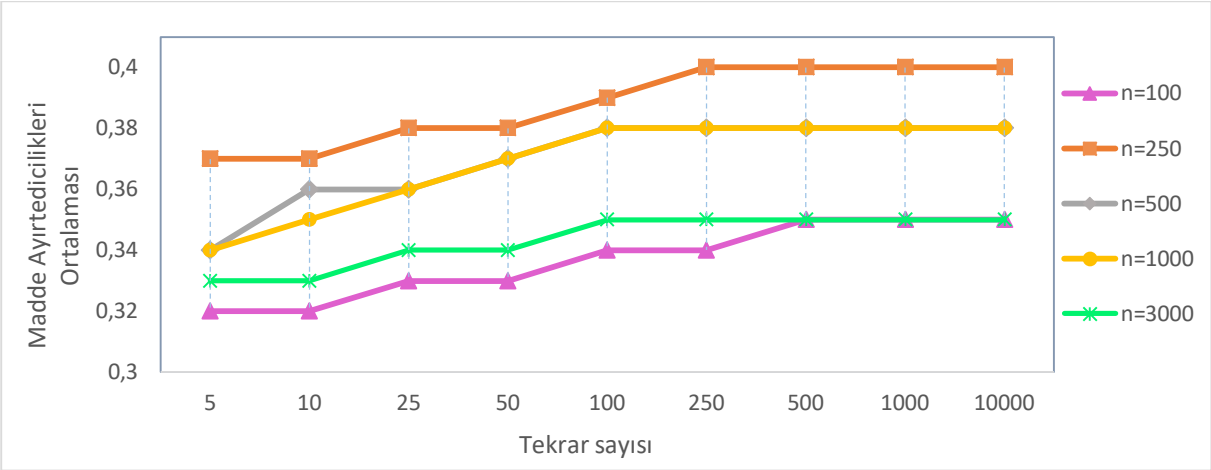
Veri üretme aşamasında madde sayısı 20, yanıt kategorisi 5'li Likert tipi ölçek olacak şekilde sabitlenmiştir. Madde parametrelerine ilişkin herhangi bir sınırlama yapılmamıştır. Her bir örneklem büyüklüğünde madde ayırt edicilikleri ortalaması farklı tekrar sayılarında hesaplanmıştır. Bu parametreye ilişkin herhangi bir kısıtlama yapılmadığı için değerlendirme yapılırken değerdeki değişimin azalıp sabitlenmeye başladığı tekrar sayısı yeterli tekrar sayısı olarak belirlenmiştir.

Tablo 3
Madde Ayırtedicilikleri Ortalaması

Tekrar sayısı	Örneklem Büyüklüğü				
	N=100	N=250	N=500	N=1000	N=3000
5	0,32	0,37	0,34	0,34	0,33
10	0,32	0,37	0,36	0,35	0,33
25	0,33	0,38	0,36	0,36	0,34
50	0,33	0,38	0,37	0,37	0,34
100	0,34	0,39	0,38	0,38	0,35
250	0,34	0,40	0,38	0,38	0,35
500	0,35	0,40	0,38	0,38	0,35
1000	0,35	0,40	0,38	0,38	0,35
10000	0,35	0,40	0,38	0,38	0,35

Tablo 3 Monte Carlo yöntemi ile üretilen veri setlerinde yer alan 20 maddenin ayırt edicilik ortalamalarını içermektedir. Şekil 3'te ise madde ayırt edicilik

ortalamaları her bir örneklem büyüklüğü için tekrar sayısına bağlı olarak grafik ile sunulmuştur.



Şekil 3. Madde Ayırt Edicilikleri Ortalaması

Şekil 3 ve Tablo 3'te yer alan madde ayırt edicilikleri ortalamaları incelendiğinde, örneklem büyüklüğü 100 olduğunda 500, örneklem büyüklüğü 250 olduğunda 250 ve örneklem büyüklüğü 500, 1000, ve 3000 olduğunda ise 100 tekrardan sonra madde ayırt edicilikleri ortalamalarının sabitlendiği görülmektedir. Buna göre örneklem büyüklüğü arttıkça gerekli tekrar sayısı azalmaktadır. Buna ek olarak tekrar sayısı arttıkça her bir örneklem büyüklüğünde madde ayırt edicilikleri ortalaması da artmaktadır ve bir noktadan sonra sabitlenmektedir.

Veri üretme aşamasında 20 maddelik tek boyutlu bir yapı belirlenerek veri üretilmiştir. DFA ile belirlenen 20 maddelik tek boyutlu yapı ve verinin uyumu incelenmiştir. Model veri uyumuna ilişkin χ2, AIC ve RMSEA değerleri kestirilmiştir. Tablo 4, Tablo 5 ve Tablo 6 model veri uyumunun tekrar sayısına bağlı değişimini içermektedir.

Tablo 4
Farklı Atama Sayılarında Model Veri Uyumunun Değişimi (χ2)

Tekrar sayısı	N=100		N=250		N=500		N=1000		N=3000	
	χ^2	$\Delta\chi^2$	χ^2	$\Delta\chi^2$	χ^2	$\Delta\chi^2$	χ^2	$\Delta\chi^2$	χ^2	$\Delta\chi^2$
5	364,80		396,07		412,35		455,82		475,39	
10	332,19	32,61*	365,61	30,46*	377,11	35,24*	432,11	23,71	444,30	31,09*
25	294,14	38,05*	332,45	33,16*	340,07	37,04*	401,79	30,32*	412,15	32,15*
50	259,37	34,77*	301,19	31,26*	306,34	33,73*	370,13	31,66*	381,66	30,49*
100	224,41	34,96*	268,56	32,63*	272,93	33,41*	332,65	37,48*	349,37	32,29*
250	220,87	3,54	261,82	6,74	239,50	33,43*	301,32	31,33*	317,93	31,44*
500	209,33	11,54	245,20	16,62	227,61	11,89	285,36	15,96	284,65	33,28*
1000	209,03	0,30	229,75	15,45	222,01	5,60	259,65	25,71	273,26	11,39
10000	207,95	1,08	227,70	2,05	219,53	2,48	257,60	2,05	271,18	2,08

*p (χ^2sd=19>30,144) <0.05

Tablo 4 χ2'yi temel alan model veri uyumunun tekrar sayısına bağlı değişimini sunmaktadır. Üretilen verilere AFA yapılarak elde edilen χ2 değerleri karşılaştırılırken bu değerler arasındaki farkın manidarlığından faydalanılmıştır. Örneğin, örneklem büyüklüğü 100 olduğunda 5 tekrar ile üretilen veriden elde edilen model veri uyumu ile 10 tekrar ile üretilen veriden elde edilen model veri uyumu karşılaştırılırken, χ2 değerleri arasındaki fark hesaplanarak manidarlığı sınanmıştır. Örneklem büyüklüğü 100 ve 250 olduğunda 100 tekrardan sonra χ2 değerinde anlamlı bir değişme olmadığı, örneklem büyüklüğü 500 ve 1000 olduğunda 250 tekrardan sonra, örneklem büyüklüğü 3000 olduğunda ise 500 tekrardan sonra χ2 değerinde anlamlı bir değişme olmadığı sonucuna ulaşılmıştır. Buna göre, örneklem büyüklüğü 100 ve 250 olduğunda daha az tekrar ile model veri uyumu sabitlenirken, örneklem büyüklüğü 500, 1000 ve 3000 olduğunda model veri uyumunun sabitleşmesi için daha fazla tekrara ihtiyaç duyulmaktadır. χ2 istatistiğinin örneklem büyüklüğüne duyarlı olduğu, küçük örneklemlerde model veri uyumu yüksek çıkarken, büyük örneklemlerde daha düşük çıktığı (Koçak, 2016) göz önünde bulundurulduğunda örneklem küçüldükçe daha az tekrar ile model veri uyumunun sağlanmasının ele alınan χ2 istatistiğinin örneklem büyüklüğüne duyarlılığından kaynaklandığı iddia edilebilir. Tablo 6'da model veri uyumunun değerlendirilmesinde başvurulan bir diğer istatistik olan RMSEA değerine ilişkin bulgular sunulmuştur.

Tablo 5
Farklı Atama Sayılarında Model Veri Uyumunun Değişimi (χ2)

Tekrar sayısı	N=100 RMSEA	N=250 RMSEA	N=500 RMSEA	N=1000 RMSEA	N=3000 RMSEA
5	0,092	,090	0,070**	0,072**	0,070**
10	0,091	,084	0,070**	0,071**	0,069**
25	0,091	,081	0,062**	0,064**	0,061**
50	0,091	,081	0,055**	0,059**	0,051**
100	0,087	,074**	0,055**	0,047*	0,046*
250	0,084	,068**	0,048*	0,044*	0,046*
500	0,080**	,057**	0,048*	0,042*	0,046*
1000	0,079**	,050*	0,047*	0,042*	0,045*
10000	0,079**	,048*	0,046*	0,041*	0,045*

*(*iyi uyum, ** kabul edilebilir uyum, Tabachnick ve Fidell, 2007)*

Tablo 6'da her bir örneklem büyüklüğü için farklı tekrar sayıları ile üretilen verilerden DFA ile elde edilen model veri uyumuna ilişkin RMSEA değerleri yer almaktadır. Tabacnick ve Fidell (2007), RMSEA değerinin 0 ile 0,05 arasında olmasının iyi uyuma, 0,05 ile 0,08 arasında olmasının ise kabul edilebilir uyuma işaret ettiğini belirtmektedir. Bu değerlendirmeye göre, örneklem büyüklüğü 100 olduğunda en az 500 tekrar ile kabul edilebilir uyum, örneklem büyüklüğü 250 olduğunda en az 100 tekrar ile kabul edilebilir uyum, 1000 tekrar ile iyi uyum düzeyine erişilmektedir. Örneklem büyüklüğü 500 olduğunda 250, örneklem büyüklüğü 1000 olduğunda 100 ve örneklem büyüklüğü 3000 olduğunda ise 100 tekrar ile iyi uyuma erişilebileceği

sonucuna ulaşılmıştır. Örneklem büyüklüğü 100 olduğunda iyi uyumun elde edilemediği, iyi uyuma erişilememesinin RMSEA değerinin küçük örneklemlerde performansının düşük olması (Tabacnick ve Fidell, 2007) ile ilişkili olduğu ifade edilebilir. Buna rağmen örneklem büyüklüğü arttıkça daha az tekrar ile daha iyi uyuma erişildiği ifade edilebilir. Örneklem büyüklüğünün değerlendirilmesinde başvurulan bir diğer uyum indeksi olan AIC değerine ilişkin kestirimler Tablo 6'da sunulmuştur.

Tablo 6
Farklı Atama Sayılarında Model Veri Uyumunun Değişimi (Model AIC)

Tekrar sayısı	N=100		N=250		N=500		N=1000		N=3000	
	AIC	ΔAIC	AIC	ΔAIC	AIC	ΔAIC	AIC	ΔAIC	AIC	ΔAIC
5	366,86		435,21		509,14		606,38		632,70	
10	364,05	2,81	434,05	1,16	506,00	3,14	604,80	1,58	628,25	4,45
25	361,96	2,09	430,87	3,18	501,84	4,16	600,13	4,67	621,72	6,53
50	355,60	6,36	422,14	8,73	496,64	5,20	597,40	2,73	619,54	2,18
100	353,38	2,22	414,38	7,76	481,47	15,17	590,28	7,12	610,43	9,11
250	349,01	4,37	411,1	3,28	476,23	5,24	583,09	7,19	607,10	3,33
500	346,75	2,26	409,95	1,15	473,50	2,73	580,00	3,09	603,76	3,34
1000	345,21	1,54	408,83	1,12	472,65	0,85	579,41	0,59	602,23	1,53
10000	343,44	1,77	407,78	1,05	471,93	0,72	578,84	0,57	601,58	0,65

Tablo 7'de AIC değerleri sunulmuştur. AIC istatistiğine bağlı model veri uyumunun değerlendirilmesinde iki farklı AIC değeri karşılaştırılarak, daha küçük olanın modele daha uyumlu olduğu yorumu yapılır (Bock ve Aitkin, 1981). Buna göre büyün örneklem büyüklüklerinde tekrar sayısı arttıkça AIC değerinin düştüğü dolayısıyla uyumun arttığı sonucuna varılmıştır.

Model veri uyumunun değerlendirilmesinde kullanılan $\chi 2$, RMSEA ve AIC değerleri birlikte değerlendirildiğinde RMSEA ve AIC değerlerinin tutarlı olduğu, örneklem büyüklüğü arttıkça model veri uyumunun sağlanması için gerekli tekrar sayısının azaldığı, küçük örneklemlerde ise model veri uyumunun sağlanabilmesi için daha çok sayıda tekrara ihtiyaç duyulduğu sonucuna ulaşılmıştır. Her bir örneklem büyüklüğü kendi içinde ele alındığında ise, tekrar sayısı arttıkça model veri uyumunun arttığı görülmüştür. $\chi 2$ istatistiğinin ise küçük örneklemlerde daha az tekrar ile, büyük örneklemlerde daha çok tekrar ile sabitlendiği bu durumun ise $\chi 2$ istatistiğinin küçük örneklemlerde model veri uyumunu yükseltici etkisinin olmasıyla ilişkili olduğu düşünülmektedir.

Elde edilen bulgulara göre, Klasik Test Kuramı'nda açıklanan varyans oranı, Cronbach Alfa katsayısı ve madde ayırt edicilikleri ortalaması ve model veri uyumunun tekrar sayısına bağlı değişimi göz önünde bulundurulduğunda, her bir parametre için gerekli en az tekrar sayısı Tablo 7'te sunulmuştur.

Tablo 7
Klasik Test Kuramı Madde ve Test Parametreleri için Monte Carlo Yönteminde Gerekli Minimum Tekrar Sayıları.

	Örneklem Büyüklüğü				
	N=100	**N=250**	**N=500**	**N=1000**	**N=3000**
Açıklanan toplam varyans oranı	1000	250	100	100	50
Cronbach Alfa katsayısı	1000	500	250	100	100
Madde Ayırtedicilikleri ortamaları	500	250	100	100	100
Model Veri Uyumu	500**	100** 1000*	250*	100*	100*

Tablo 7 incelendiğinde, araştırma sonucunda örneklem büyüklüğü 100 olduğunda açıklanan toplam varyans oranı ve Cronbach Alfa değerleri için minimum 1000 tekrara, madde ayırt edicilikleri ortalaması ve model veri uyumu indeksleri için ise minimum 500 tekrara ihtiyaç duyulduğu görülmektedir. Üretilen yapay bir veride bu değerlerin tümünün önemli olduğu düşünüldüğünde örneklem büyüklüğü 100 olduğunda minimum 1000 tekrara ihtiyaç duyulduğu ifade edilebilir. Benzer değerlendirme örneklem büyüklüğü 250 olan koşul için yapıldığında bu koşulda gerekli minimum tekrar sayısının 500 olduğu, bu tekrar sayısı ile üretilecek verilerde kabul edilebilir model veri uyumu sağlanacaktır. Örneklem büyüklüğü 500 olduğunda minimum 250, örneklem büyüklüğü 1000 ve 3000 olduğunda ise minimum 100 tekrara ihtiyaç olduğu ifade edilebilir.

Tartışma, Sonuç ve Öneriler

Bu araştırmada Monte Carlo yöntemi kullanılarak üretilen verilerde tekrar sayısının KTK test ve madde parametrelerine etkisinin belirlenmesi ve Monte Carlo yöntemi ile veri simülasyonu yapılırken hangi koşulda en az kaç tekrara ihtiyaç olduğunun belirlenmesi amaçlanmıştır. Bu amaçla verilerde madde sayısı, yanıt kategorisi ve yapı sabitlenerek 20 maddelik, çoklu puanlanan, tek boyutlu olacak şekilde örneklem büyüklüğü 100, 250, 500, 1000 ve 3000 olan veriler tekrar sayısı 5, 10, 25, 50, 100, 250, 500, 1000 ve 10000 olarak değişimlenerek üretilmiştir. Üretilen verilerde, açıklanan toplam varyans oranı %40, Cronbach Alfa iç tutarlılık katsayısı 0.70 olarak sınırlandırılmış, model veri uyumu ve madde ayırt edicilik ortalamalarına herhangi bir kısıtlama getirilmemiştir.

Araştırma sonucunda her dört parametre için de $\chi 2$ istatistiği göz ardı edildiğinde örneklem büyüklüğü arttıkça gerekli tekrar sayısının azaldığı sonucuna ulaşılmıştır. Binois, Huang, Gramary ve Ludkovsk (2019), Gifford ve Swaminathan (1990) ve Harwell, Rubinstein, Hayes ve Olds, (1992), örneklem büyüklüğü arttıkça gerekli tekrar sayısının azalacağını, Sobol (1971) ve Yaşa (1996) Monte Carlo yönteminde tekrar sayısı artırıldığında örneklemdeki değişkenliğin düşüreceğini yani sabit değerlere erişileceğini, Brooks (2002) ve Hutchinson ve Bandalos (1997) yetersiz tekrar sayısının yanlış tahminlere yol açacağını ifade etmektedir. Araştırma sonuçları bu ifadeleri destekler niteliktedir.

$\chi 2$ İstatistiği diğer uyum indekslerinin aksine küçük örneklemlerde daha az sayıda tekrar ile sabitleşmiştir. Hambleton ve diğerleri (1991), $\chi 2$ dağılımının örneklem büyüklüğüne karşı oldukça hassas olduğunu, büyük örneklemler için model veri

uyumunun bu istatistikle neredeyse hiç sağlanamadığını belirtmektedir. Örneklemin küçük olması $\chi 2$ istatistiğine dayalı uyum iyiliğini artırmaktadır (Bock, 1997; Drasgow ve ark., 1995) buna göre örneklem büyüklüğü küçük olan koşullarda uyumun daha az tekrarla sabitlenmesinin $\chi 2$ istatistiğinin küçük örneklemlerde daha uyumlu sonuç vermesi ile ilgili olduğu söylenebilir. AIC ve RMSEA istatistiklerinde ise örneklem büyüklüğü arttıkça daha az tekrar sayısı ile daha iyi uyuma erişilmektedir.

Cronbach Alfa kestiriminde örneklem büyüklüğü arttıkça gerekli tekrar sayısı azalmıştır. Belirlenen 0.70 değerine örneklem büyüklüğü 100 olduğunda 1000 tekrar ile örneklem büyüklüğü 250 olduğunda ise 500 tekrar ile erişilmiştir. Alanyazında Cronbach Alfa'nın yansız kestirilebilmesi için minimum örneklem büyüklüğü ile ilgili 100 (Yurdugül, 2008) ve 300 (Kline, 1986), 400 (Charter, 1999) olması gerektiği, 300'ün yetersiz olacağı (Nunnally ve Berstein, 1994; Segall, 1994) gibi yer almaktadır. Bu ifadeler göz önünde bulundurulduğunda 100 ve 250 örneklem büyüklüklerinde Cronbach Alfa kestiriminde hatanın daha yüksek olabileceği ifade edilebilir. Bu nedenle, küçük örneklemde karşılaşılan yanlılığın giderilebilmesi ve hedeflenen düzeyde güvenirliğe erişilebilmesi için Monte Carlo yönteminde daha fazla tekrara ihtiyaç duyulmaktadır. Açıklanan toplam varyans oranının elde edildiği Açımlayıcı Faktör Analizi için de benzer durum geçerlidir. Alanyazında EFA için gerekli örneklem büyüklüğünün 200 (Büyüköztürk, 2002) olması gerektiği ya da madde sayısının en az 10 katı (Kline, 1986) yani bu örnek için 200 ve 5 (Child, 2006) katı yani bu örnek için 100 olması gerektiği yönünde ifadeler yer almaktadır. Buna göre 100 ve 250 örneklem büyüklüklerinin faktörleştirme için sınırda olduğu ve az olduğu ifade edilebilir. Dolayısıyla yapılan analiz için gerekli örneklem büyüklüğü sağlanmadığı koşulda Monte Carlo yöntemi için gerekli tekrar sayısı daha fazla olacaktır. Bu nedenle üretilen veride belirlenen koşullar yapılacak analizlerin varsayımını karşılamıyorsa eğer araştırmacıların daha fazla tekrar ile veri üretmesi önerilebilir.

Açıklanan toplam varyans oranında ve Cronbach Alfa kestiriminde örneklem büyüklüğü 100 olduğunda, 1000 tekrar ile üretme aşamasında belirlenen değerlere erişildiği görülmektedir. Aynı örneklem büyüklüğünde madde ayırt edicilikleri ortalaması ve model veri uyumunda ise daha az tekrar sayısı ile sabit değerlere erişilmiştir. Buna göre küçük örneklemlerde kısıtlama getirilen yani manipüle edilen parametrelerde daha fazla tekrar gerekirken, serbest bırakılan parametrelerde daha az sayıda tekrar ile veri üretilebilmektedir. Örneklem küçük olduğunda simülasyon ile manipüle edilen aralıklara ulaşılması güçtür (Baker, 1998; Brown, 1994; Goldman ve Raju, 1986; Hulin, Lissak ve Drasgow, 1982; Lord, 1968), ve simülasyon aşamasında parametreye aralık tanımlanıyorsa daha fazla tekrara ihtiyaç duyulacaktır (Brooks, 2002; Hutchinson ve Bandalos; 1997; Sobol, 1971) bu nedenle örneklem büyüklüğü 100 olduğunda açıklanan varyans oranı ve Cronbach Alfa için daha fazla tekrara ihtiyaç duyulduğu ifade edilebilir. Bir diğer ifadeyle, her hangi bir parametreye ilişkin veri üretme aşamasında manüplasyon yapılıyorsa, istenilen özelliği taşıyan veri setinin Monte Carlo yöntemi ile elde edilebilmesi için daha fazla tekrara ihtiyaç duyulacaktır.

Alanyazında Monte Carlo yöntemi ile Klasik Test Kuramına ilişkin çalışmalar bulunmasına karşın bu çalışmalarda tekrar sayısının kaç olması gerektiğine ilişkin bir bulgu yer almamaktadır. Mundform ve diğerleri (2011), genel olarak simülasyon çalışmalarında tekrar sayısını belirlemede bir prosedürün olmadığını bu nedenle tekrar sayısının araştırmacının insiyatifinde olduğunu ifade etmektedir. Yapılan

Monte Carlo simülasyon çalışmalarda baş vurulan tekrar sayısının farklı olduğu ve belirlenen bu tekrar sayılarının araştırmacının kanaati ile olduğu ve bir gerekçeye dayandırılmadığı görülmektedir. Harwell ve diğerleri (1992) ise kaç tekrara başvurulması gerektiğini söylemenin koşullara ve analize bağlı olduğunu belirtmektedir. Stone (1993), bağımsız değişkenin alacağı değerlere de bağlı olmak şartıyla tekrar sayısının en az 25 olması gerektiğini, bu durumda MC yönteminin gücünü artıracağını belirtmiştir.

Bu çalışmada tek boyutlu bir yapıda madde sayısı 20, yanıt kategorisi 5 olarak sabitlenmiş ve örneklem büyüklüğü 100, 250, 500, 1000 ve 3000 olarak değişimlenerek, açıklanan toplam varyan oranı, Cronbach Alfa, madde ayırt edicilikleri ortalaması ve model veri uyumu parametrelerinin kestimini için Monte Carlo yönteminde gerekli tekrar sayısının kaç olduğu sorusuna yanıt aranmıştır. CTT temel alınarak yapılacak bir çalışmada araştırmacılara örneklem büyüklüğü 100 olduğunda 1000, örneklem büyüklüğü 250 olduğunda 500, örneklem büyüklüğü 500 olduğunda 250 ve örneklem büyüklüğü 1000 ve 3000 olduğunda 100 tekrar ile veri üretmeleri önerilmektedir.

Kaynakça

Aiken, L, R. (2000). *Psychological testing and assessment*. Boston. Allyn and Bacon.

Baker, F. B. (1998). An investigation of the item parameter recovery of a Gibbs sampling procedure. *Applied Psychological Measurement, 22*, 153-169. https://doi.org/10.1177/01466216980222005

Binois M., Huang J., Gramacy R.B., and Ludkovski M. (2019). Replication or exploration? Sequential design for stochastic simulation experiments. https://arxiv.org/abs/1710.03206. DOI: 10.1080/00401706.2018.1469433

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

Bock, R.D. (1997). A brief history of item response theory. *Educational Measurement: Issues and Practice.* Winter 1997.

Brooks, C. (2002). *Introductory econometrics for finance*. Cambridge University Press.

Brown, R. L. (1994). Efficacy of the indirect approach for estimating structural equation models with missing data: A Comparison of Methods. *Structural Equation Modeling:A Multidisciplinary Journal. 1*(4), 287-316. https://doi.org/10.1080/10705519409539983

Büyüköztürk, Ş. (2002). *Sosyal bilimler için veri analizi el kitabı.* Ankara: Pegem Yayıncılık.

Charter, R.A. (2003). Study samples are too small to produce sufficiently precise reliability coefficients. *The Journal of General Psychology, 130,* 117-129. https://doi.org/10.1080/00221300309601280

Child, D. (2006). *The Essentials of factor analysis.* Continuum, London.

Çelen, Ü. (2008). Klasik Test Kuramı ve Madde Tepki Kuramına dayalı olarak geliştirilen iki testin psikometrik özelliklerinin karşılaştırılması. *İlköğretim Online, 7*(3),758-768.

De Ayala, R. J. (2009). *The theory and practice of item response theory*. New York,

Dooley, K. (2002). *Simulation research methods.* In J. Baum (Ed.). Companion to organizations. London: Blackwell.

Drasgow, F., Levine, M., Tsien, S., Williams, B., and Mead, A. (1995). Fitting polytomous item response theory models to multiple-choice tests. *Applied Psychological Measurement, 19(*2), 143-165. https://doi.org/10.1177/014662169501900203

Fay D.S., and Gerow K. A (2013). *Biologist's guide to statistical thinking and analysis* , WormBook, ed. The C. Elegans Research Community, WormBook, doi/10.1895/wormbook.1.159.1, http://www.wormbook.org. https://doi.org/10.1895/wormbook.1.159.1

Gifford, J.A., and Swaminathan, H. (1990), Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement 27*(4), 361-370. https://doi.org/10.1111/j.1745-3984.1990.tb00754.x

Glen Satten A., Flanders W. D., and Yang Q. (2001). Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. *Am. J. Hum. Genet. 68*:466–477. https://doi.org/10.1086/318195

Goldman, S.H., and Raju, N. S. (1986). Recovery of one- and two-parameter logistic item parameters: An empirical study. *Educational and Psychological Measurement, 46*(1), 11-21. https://doi.org/10.1177/0013164486461002

Hambleton R.K., Swaminathan H. and H. J. Rogers (1991). *Fundamentals of item response theory.* Newbury Park, CA: SAGE Publications, Inc.

Hambleton, R. K., Jones R.W., and Rogers, H. J. (1993). Influence of item parameter estimation errors in test development. *Journal of Educational Measurement, 30,* 143-155. https://doi.org/10.1111/j.1745-3984.1993.tb01071.x

Hammersley, J.M., and Handscomb, D.C. (1964). *Monte-Carlo Methods.* Springer Netherlands http://dx.doi.org/10.1007/978-94-009-5819-7

Harwell, M. R., and Janosky, J. E. (1991). An empirical study of the effects of small datasets and varying prior distribution variances on item parameter estimation in BILOG. *Applied Psychological Measurement*, *15*, 279-291. https://doi.org/10.1177/014662169101500308

Harwell, M. R., Stone, C. A., Hsu, T. C., and Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement*, *20*(2), 101-125. https://doi.org/10.1177/014662169602000201

Harwell, M.R., Rubinstein E., Hayes W.S., and Olds, C. (1992). Summarizing Monte Carlo results in methodological research: The fixed effects single- and two-factor ANOVA cases. *Journal of Educational Statistic, 17*, 315-339. https://doi.org/10.3102/10769986017004315

Hauck, W.W., and Anderson, S. (1984). A new statistical procedure for testing equivalence in two-group comparative bioavailability trials. *Journal of Pharmacokinetics and Biopharmaceutics, 12*, 83-91. https://doi.org/10.1007/BF01063612

Hulin, C. L., Lissak, R. I., and Drasgow, F. (1982). Recovery of two and three parameter logistic item characteristic curves: A Monte Carlo study. *Applied Psychological Measurement, 6,* 249-260. https://doi.org/10.1177/014662168200600301

Hutchinson, S.R., and Bandalos, D.L. (1997). A guide to Monte Carlo simulations for applied researchers. *Journal of Vocational Education Research*, 22(4), 233-245.

Kannan P., Sgammato A., Tannenbaum R.J., and Katz I.R. (2015) Evaluating the consistency of Angoff-Based cut scores using subsets of items within a generalizability theory framework, *Applied Measurement in Education*, *28*(3), 169-186. https://doi.org/10.1080/08957347.2015.1042156

Kéry, M., and Royle J. A. (2016). *Applied hierarchical modeling in ecology* Volume 1: Prelude and Static Models Book.ScienceDirect.

Kline, P. (1986) *A handbook of test construction: Introduction to psychometric desing. New York:* Methune and Company.

Koçak, D. (2016). *Kayıp veriyle baş etme yöntemlerinin Madde Tepki Kuramı bir parametereli lojistik modelinde model veri uyumuna ve standart hataya etkisi*. (Yayımlanmamış doktora tezi), Ankara Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara.

Leventhall, B., and Ames, A. (2019). *Using SAS for Monte Carlo Simulation Studies in Item Response Theory.* National Council on Measurement in Education Annual Meeting in Toronto, Ontorio Canada.

Lewis, P.A.W., and E.J. Orav. (1989). *Simulation methodology for statisticians, operations analysts, and engineers* . Volume 1. Wadsworth and Brooks/Cole, California: Pacific Grove.

Lord, F. M., and Novick, M. R.(1968*). Statistical theories of mental test scores*. Reading MA: Addison- Wesley.

Mundform, D.J., Schaffer, J., Myoung-Jin, K., Dale, S, Ampai, T., and Pornsin S.(2011). Number of replications required in Monte Carlo simulation studies: A synthesis of four studies. *Journal of Modern Applied Statistical Methods: 10(1),* Article 4. Available at: http://digitalcommons.wayne.edu/jmasm/vol10/iss1/4

Murie C., and Nadon R. (2018). A correction for the LPE statistical test. Bioconductor. https://www.bioconductor.org/packages/devel/bioc/vignettes/LPEadj/inst/doc/LPEadj.pdf

Naylor, T.H.,Blantify J., Burdick D.S., and Chu K. (1968). *Cumputer simulation techniques*. John Wiley and Sons, New York.

Nunnally, J.C., and Bernstein, J.H. (1994). *Psychometric theory.* New York: McGraw-Hill. New Y: The Guilford Press.

Qualls, A. L., and Ansley, T. N. (1985, April). A comparison of item and ability parameter estimates derived from LOGIST and BILOG. Paper presented at the meeting of the *National Council on Measurement in Education*, Chicago.

R Development Core Team (2011), *R: A language and environment for statistical computing, a foundation for statistical computing,*.Vienna, Austria, ISBN 3900051-070, Erişim:[http://www.R-project.org].

Rubinstein, R. Y. (1981). *Simulation and the Monte Carlo method.* John Wiley and Sons, New York. https://doi.org/10.1002/9780470316511

Saeki H., and Tango T. (2014) Statistical inference for non-inferiority of a diagnostic procedure compared to an alternative procedure, based on the difference in correlated proportions from multiple raters. In: van Montfort K., Oud J., Ghidey W. (eds) *Developments in Statistical Evaluation of Clinical Trials.* Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-55345-5_7

Segall, D.O. (1994). The reliability of linearly equated tests. *Psychometrika, 59*, 361-375. https://doi.org/10.1007/BF02296129

Sobol I.M. (1971). *The Monthe Carlo method*. Moscow, Russian.

Stone, C. A. (1993). The use of multiple replications in IRT based Monte Carlo research. Paper presented at the *European Meeting of the Psychometric Society*, Barcelona.

Tabachnick, B. G., and Fidel, L. S. (1996). *Using multivariate statistics*. (3. Ed). MA:AllynandBacon, Inc.

Yaşa, F. (1996). *Rasgele değişen bazı fiziksel olayların 3 boyutlu monte carlo yöntemi ile modellenmesi.* (Yayınlanmamış yüksek lisans tezi). Kahramanmaraş Sütçü İmam Üniversitesi / Fen Bilimleri Enstitüsü. Kahramanmaraş,Türkiye.

Yen, W. M. (1987). A comparison of the efficiency and accuracy of BILOG and LOGIST. *Psychometrika, 52*, 275-291. https://doi.org/10.1007/BF02294241

Yo, H., Han, K.T., and Oh, H.J. (2019). *Software Packages for Item Response Theory–Based Test Simulation: WinGen3, SimulCAT, MSTGen, and IRTEQ*. National Council on Measurement in Education Annual Meeting in Toronto, Ontorio Canada.

Yurdugül, H. (2008). Cronbach alfa katsayısı için minimum örneklem genişliği: Monte Carlo çalışması. *H.U. Journal of Education, 35,* 397-405.

Summary

Introduction

In the field of education and psychology, the characteristics that are subject to measurement are often not directly observed and therefore other features that are considered to be indicative of the characteristics are utilized. For example, since abstract features that are cognitive and affective cannot be observed directly, they are often measured by means of a test or scale. Based on the individual's responses to the items that make up the test, an inference is made about the relevant feature. The test developed in order to measure a feature and the data to be obtained as a result of this test should be based on a test theory. It can be said that it is the most preferred test theory since the assumptions of Classical Test Theory (CTT) are easily fulfilled.

Although Classical Test Theory is the most widely used test theory, it has some disadvantages. These disadvantages bring about questions about conditions that will eliminate or minimize these disadvantages. For example, when the data obtained as a result of applying a scale to a particular group has a high ratio missing value that is at the missing completely at random mechanism, which missing value method makes predictions with high realibitiy and validity?

Sobol (1971) states that increasing the number of repetitions of the Monte Carlo study is necessary to provide truthful results. However, no study to date has suggested how much replication is enough for validity and reliability. In addition, while determining the number of repetitions used in researches, it is not stated how the replication number was decided. Binois, Huang, Gramary and Ludkovsk (2019) state that the number of repetitions required in the Monte Carlo method gets smaller as the sample grows, and this is a common point of all simulation methods. Accordingly, it can be stated that in small samples are needed more and large samples require less replication (Harwell, Rubinstein, Hayes and Olds, 1992). In the literature,

there are studies using different repetition numbers such as 10000 (Saeki ve Tango, 2014), 1000 (Kannan, Sgammato, Tannenbaum and Katz 2015; Bionis, Huang and Gramacy, 2019), 500 (Glen Satten, Flanders and Yang, 2001) 100 (Fay and Gerow; 2013), 10 (Kéry and Royle, 2016; Murie and Nadon, 2018). However, there are also simulation studies in which no replications are performed or the number of replications is not reported (Hambleton, Jones and Rogers, 1993; Harwell and Janosky, 1991; Hulin, Lissak and Drasgow, 1982; Qualls and Ansley, 1985; Yen, 1987).

The importance of the number of repetitions in the simulation studies to produce truth-reflecting results is indisputable. When a research is designed using Monte Carlo simulation technique, the number of repetitions is very important for the reliability and validity of the research results. However, there is no clear information on how many repetitions are sufficient. In this study, it is aimed to determine the effect of number of repetitions in Monte Carlo simulation method on item and test parameter estimations in Classical Test Theory and to determine the number of repetitions required. For this purpose, the data obtained by changing the number of replication under different conditions total variance ratio Cronbach's alpha coefficient average of item discrimination and model-data-fit parameters were examined.

Method

In this study, it was aimed to determine the effect of replication number on test and item parameters and required replication number in data produced by Monte Carlo simulation method based on CTT. This study is a Monte Carlo simulation study.

In the research, R program (2011) "psych" package was used for data generation and analysis. Monte Carlo Simulation steps are presented below:

Table 1
Monte Carlo simulation steps

Process	Operation
1. Research questions that reflect the purpose of the study.	What is the effect of Replication number on Reliability Validity in the Monte Carlo CTT simulation studies? How much replications are sufficient in the Monte Carlo CTT simulation studies?
2. Variables and conditions (levels)	• Sample size: 100, 250, 500, 1000 and 3000 • Test length: 20 • response category: 5 (1-0) • Replication number: 5, 10, 25, 50, 100, 250, 500, 1000 and 10000 • *Total eigenvalue: %40 • *Cronbach's Alpha coefficient: 0,70 • Dimensional of data: One In the Data that was generated the data dimension was restricted as one, total eigenvalue was restricted as 40%, and Cronbach's Alpha coefficient was restricted as 70.
3. Creating appropriate experimental design.	Number of sample size conditions x number of test length conditions x number of replication numbers conditions= 5x1x9= 45 different simulation conditions.
4. Data generation.	Experimental scenarios with the specified conditions are simulated based on CTT.

5. Estimation of parameters (Analysis)	• Reliability: Cronbach's alpha coefficient was calculated as reliability coefficient. The obtained Cronbach Alpha coefficient values were interpreted based on the 0.70 value determined during the data generation stage. • Validity: The total eigenvalue was calculated as evidence of validity In the Exploratory Factor Analysis. The total variances obtained by exploratory factor analysis were interpreted based on the value of 40% determined in the data generation stage. The model-data fit was calculated as evidence of the validity in the confirmatory factor analysis. One-dimensional structure was tested by confirmatory factor analysis and model data fit was estimated. $\chi 2$, AIC, and RMSEA values were used to evaluate model data fit. In the comparison of $\chi 2$ data fit statistics, the significance of the difference between $\chi 2$ values was tested. In the evaluation of AIC statistic, it is evaluated that the model with smaller value is fit with the data. For this reason, descriptive comparison method was used to compare the AIC values estimated from the data obtained from different replications of the same sample size. The interpretation of the RMSEA value is based on the value ranges defined by Tabachnick and Fidell (2007). According to this, $0 <$ good fit $\leq 0,05$ and $0,05 <$ acceptable fit $\leq 0,08$ was used as model-data fit index. Average of item discrimination was calculated as evidence of the validity. The mean of item discrimination was calculated by calculating the index of discrimination of the items.
6. Comparison.	The validity and reliability indices of the data sets were compared according to different number of repetitions under the same condition. Cronbach's Alpha coefficient, explained variance ratio and mean of item discriminant were compared descriptively. When searching for answers to the question of how many repetitions are sufficient, the number of repetitions reached to the values determined (0.70 for Cronbach's Alpha coefficient, 40% for explained variance ratio) in the data generation stage was determined. The significance of the difference between $\chi 2$ values was tested while evaluating model data fit. AIC and RMSEA values were interpreted descriptively. In terms of model data fit and item discrimination mean, the number of repetitions required was interpreted considering how many repetitions these values were fixed after.
7. Repeat process for all conditions.	Each condition was compared by making calculations separately.
8. Evaluation of analysis results	The results were evaluated within the framework of the research questions.

Results and Discussion

Results of the research about total variance ratio show that the reference value is reached with 1000 repetitions when the sample size is 100, with 250 repetitions when the sample size is 250, with 100 repetitions when sample size are 500 and 1000, and with 50 repetitions when the sample size 3000. It is seen that the predictions obtained after these repeat numbers are fixed at the specified reference value. As the sample size increased, it was concluded that the total variance ratio determined in the data generation stage was reached with fewer replications.

The reference value of Cronbach's Alpha coefficient which was determined in the data generation stage is reached with 1000 replications when the sample size is 100,

with 500 replications when the sample size is 250, with 250 replications when sample size are 500, with 100 replications when the sample size are 1000 and 3000. Accordingly, as the sample size increases, the number of replication required to reach the value determined in the data generation stage decreases. Therefore, if there is a limitation on the internal consistency of the data produced on the basis of CTT, the sample size should also be taken into consideration. The number of replications required for Monte Carlo method decreases when the sample size is increasing. The test length was fixed as 20, and response category of items was fixed as a five-category Likert scale in the data generate stage. Item parameters were not limited. The mean of item discrimination was calculated in each sample size for each replicate number. Since there are no restrictions on this parameter, the number of repetitions in which the change in value decreases and starts to be fixed was determined as sufficient repetitions.

When the item discrimination averages were examined, it was seen that when the sample size is 100 after the 500 replications, when the sample size is 250 after the 250 replication numbers, when the sample size is 500, 1000 and 3000 after the 100 replications number the average of the item discrimination is being fixed. Accordingly, while the sample size increases, the number of repetitions required decreases. In other words, in simulation studies with Monte Carlo method, as the sample size increases, the standard result is reached with less repetitions. In addition, as the number of repetitions increases, the mean of item discrimination in each sample size increases and is fixed after a point.

When the sample size is 100 and 250, the model data fit is fixed with less replications, whereas when the sample size is 500, 1000 and 3000, more replication is needed to stabilize the model data fit for $\chi 2$. When the $\chi 2$, RMSEA and AIC values used in the evaluation of the model data fit were evaluated together, it was concluded that the RMSEA and AIC values were consistent, the number of replications required to ensure the model data consistency decreased as the sample size increased, and more replications were needed to ensure the model data fit in the small samples. When each sample size is taken into consideration, it is seen that the model data fit increases as the number of replication increases.

Pedagogical Implications

In this study, the number of items in a one-dimensional structure is fixed to 20, the response category is 5, and the sample size is changed to 100, 250, 500, 1000 and 3000, and the total variance ratio, Cronbach Alpha, is explained for the estimation of item mean and model data fit parameters. In the Carlo simulation method, the number of replications required was answered. In a study based on CTT, it is suggested that researchers produce data with 1000 replications when sample size is 100, 500 replications when sample size is 250, 250 replications when sample size is 500 and 100 replications when sample size is 1000 and 3000. A similar study can be conducted based on other test theories.

Araştırmanın Etik Taahhüt Metni

Yapılan bu çalışmada bilimsel, etik ve alıntı kurallarına uyulduğu; toplanan veriler üzerinde herhangi bir tahrifatın yapılmadığı, karşılaşılacak tüm etik ihlallerde "Cumhuriyet Uluslararası Eğitim Dergisi ve Editörünün" hiçbir sorumluluğunun

olmadığı, tüm sorumluluğun Sorumlu Yazara ait olduğu ve bu çalışmanın herhangi başka bir akademik yayın ortamına değerlendirme için gönderilmemiş olduğu sorumlu yazar tarafından taahhüt edilmiştir.

Authors' Biodata/ Yazar Bilgileri

Duygu KOÇAK Alanya Alaaddin Keykubat Üniversitesi, Eğitim Fakültesi, Eğitim Bilimleri Bölümü'nde Dr. Öğr. Üyesi olarak çalışmaktadır.

Duygu Koçak is working as an an Assistant Professor at Alanya Alaaddin Keykubat University, Faculty of Education, Department of Educational Sciences.