Ahi Evran Üniversitesi
Kırşehir Eğitim Fakültesi Dergisi

ISSN: 2147 - 1037

http://kefad.ahievran.edu.tr

A Psychometric Comparison of Sato Test Theory with Classical Test Theory and Item Response Theory

Sait Çüm
Selahattin Gelbal

Article Information

DOI: 10.29299/kefad.902992

Received: 25.03.2021
Revised: 15.11.2021
Accepted: 03.02.2022

Keywords:

Sato Test Theory,
Student-Problem Chart Analysis,
Classical Test Theory,
Item Response Theory,
Test Development,
Psychometrics

Abstract

The aim of this study was to comparing the psychometric properties of the mathematics subtest items of the Determination of Students' Achievement Exam and the achievement levels of the students who took the mathematic subtest with Sato Test Theory (STT), Classical Test Theory (CTT) and Item Response Theory (IRT) indicators. The research was conducted on 15461 8th grade students who participated in the exam in 2005. The data of this study were analyzed by examining the correlations between item discriminations, item difficulties and individual characteristics calculated in the context of different test theories. In addition, problematic test items were analyzed by clustering and observing common elements. After the analysis, it was seen that Sato Test Theory produced similar results with other theories many times in terms of determining item and individual characteristics. Moreover, some advantages related to theory were also suggested in the study. The results obtained support the claims in the literature that STT can be considered as an alternative test theory that can allow valid and reliable measurements with predictions that robust and do not contradict with other test theories.

Sato Test Kuramı'nın Klasik Test Kuramı ve Madde Tepki Kuramı ile Psikometrik Açıdan Karşılaştırılması

Makale Bilgileri

DOI: 10.29299/kefad.902992

Yükleme: 25.03.2021
Düzeltme: 15.11.2021
Kabul: 03.02.2022

Anahtar Kelimeler:

Sato Test Teorisi,
Öğrenci-Madde Çizelgesi,
Klasik Test Kuramı,
Madde Tepki Kuramı,
Test Geliştirme,
Psikometri

Öz

Bu araştırmanın amacı, Öğrenci Başarılarının Belirlenmesi Sınavı'nın (ÖBBS) matematik alt testi maddelerinin psikometrik özelliklerinin ve testi alan öğrencilerin performans/başarı/yetenek düzeylerinin Sato Test Kuramı (STK), Klasik Test Kuramı (KTK) ve Madde Tepki Kuramı (MTK) ile belirlenmesi ve elde edilen bulguların karşılaştırılarak incelenmesidir. ÖBBS'ye (2005) katılan 15461 8.sınıf öğrencisi üzerinde yürütülen bu araştırmanın verilerifarklı test kuramları bağlamında hesaplanan madde ayırıcılıkları, madde güçlükleri ve birey özellikleri arasındaki korelasyonların incelenmesi, bunun yanı sıra sorunlu maddelerin kümelenerek ortak elemanlarının gözlemlenmesi şeklinde analiz edilmiştir. Yapılan analizler sonrasında STK'nın madde ve birey özelliklerinin belirlenmesi bakımından diğer kuramlarla pek çok kezbenzer sonuçlar ortaya koyduğu görülmüştür. Bununla birlikte çalışmada, kurama ilişkin bazı avantajlar da öne sürülmüştür. Ulaşılan sonuçlar, alanyazında yer alan, STK'nın, diğer test kuramlarıyla çelişmeyen kestirimleriyle geçerli ve güvenilir ölçmeler yapılmasına olanak tanıyabilecek alternatif bir test kuramı olarak değerlendirilebileceği iddialarını desteklemektedir.

Sorumlu Yazar: Sait Çüm, Dr, Milli Eğitim Bakanlığı, Türkiye, saitcum@hotmail.com, ORCID ID: 0000.0002.0428.5088

Yazar2: Selahattin Gelbal, Prof. Dr., Hacettepe Üniversitesi, Türkiye, sgelbal@gmail.com, ORCID ID: 0000.0001.5181.7262

Atıf için: Çüm, S., & Gelbal, S. (2022). Sato Test Kuramı'nın klasik test kuramı ve madde tepki kuramı ile psikometrik açıdan karşılaştırılması. *Kırşehir Eğitim Fakültesi Dergisi*, *23*(2), 1797-1829.

Giriş

Eğitimde testler, seçme, hazırbulunuşluğu belirleme, yönlendirme, durum belirleme, düzey belirleme, araştırma gibi çeşitli amaçlarla kullanılabilirler. Bu testlerin geliştirilmesi ve test sonuçlarının analiz edilmesi söz konusu olduğunda sıklıkla kullanılan iki önemli test kuramından söz edilebilir. Bunlar, Klasik Test Kuramı (KTK) ve Madde Tepki Kuramı (MTK) olarak adlandırılmaktadır. Alanyazında her iki kuramın da birbirlerine göre üstün ve zayıf yönlerinin olduğu ileri sürülmektedir. MTK, içerdiği ileri düzey istatistiksel yöntemlerin varsayımlarından kaynaklı olarak büyük (geniş) örneklemlere ihtiyaç duymaktadır (de Ayala, 2009; van der Linden ve Hambleton, 1997; Hulin, Lissak ve Drasgow, 1982; Ree ve Jensen, 1983). Bu nedenle, küçük gruplarla yürütülen ölçme ve değerlendirme çalışmalarında KTK'nın kullanımı yaygın bir şekilde devam etmektedir. Bununla birlikte, KTK'ya dayalı olarak hesaplanan madde istatistiklerinin gruba bağlı olması ve bireylerin başarı/yetenek puanlarının da madde/test özelliklerine bağlı olması pratikte pek çok sorunu beraberinde getirmektedir. KTK ile yapılan kestirimlerin tutarlılığı da tartışma konusudur. Çüm, Gelbal ve Tsai (2016), KTK'ya dayalı olarak elde edilen madde istatistiklerinin aynı evrene ait farklı örneklemler (30'ar kişilik) arasında büyük farklılıklar gösterdiği bulgusuna ulaşmışlardır. Bu noktada, küçük örneklemlerde daha etkili ölçmeler yapabilmek ve KTK'da karşılaşılan sorunları aşabilmek bakımından yeni test kuramları üzerinde araştırmalar yapılması ihtiyacı doğmaktadır.

1970'lerde Japon araştırmacı Takahiro Sato tarafından Öğrenci-Madde Çizelgesi Analizi (Student-Problem Chart Analysis) adında yeni bir teknik oluşturulmuştur. Analiz sonucunda elde edilen katsayıların, öğretmenlerin öğrencilerin performanslarını ve test maddelerini formüle ederek tanılamaları (diagnose), sonrasında ise öğrencilere rehberlik edilmesi ve öğretimin geliştirilmesi için referans görevi görebileceği belirtilmiştir. Bu çizelgenin kullanılmasının, az öğrencili sınıfların biçimlendirici değerlendirmesi için son derece iyi ve etkili bir yaklaşım olduğu görüşü öne sürülmüştür (Takeya, 1980; Tatsuoka, 1984).

Öğrenci- Madde Çizelgesi Analizi'nin (Ö-MÇA)2010 yılında Nagai'nın önerisiyle Gri İlişki Analizi (Grey Relational Analysis) ile güçlendirilmesi öğrencilerin başarı/yetenek düzeylerinin elde edilebilmesini de mümkün hale getirmiştir. Gri İlişki Analizi Ju-Long Deng tarafından 1982 yılında ortaya konulan Gri Sistem Teorisi (Grey System Theory) içerisinde yer alan analiz yöntemidir. Bir derecelendirme, sınıflandırma ve karar verme tekniği olan Gri İlişki Analizi, normal dağılım varsayımını sağlayamayan, belirsizlik nedeniyle modellenemeyen ve yeterli veri içermeyen kısıtlı verilerle karar verme durumlarında istatistiksel çözümler önermektedir (Deng, 1982). Bu durum az sayıda kişi/öğrenci üzerinde yapılan ölçmelerde artık "Sato Test Kuramı"-STK- (Sato Test Theory) olarak anılmaya başlayan yaklaşımı avantajlı hale getirmektedir. Günümüzde STK bağlamında, Gri Ö-MÇA (Grey S-P Chart Analysis), Rasch Gri Ö-MÇA (Rasch Grey S-P Chart Analysis)ve Kısmi Kredili Ö-MÇA (Partial Credit S-P Chart Analaysis) gibi modellemeler ile hem ikili puanlanan hem de

çoklu puanlanan madde türlerine uygulanabilen, test geliştirme ve uygulama alanında mevcut sorunlara çözüm getirebilecek çalışmalar üzerinde durulmaktadır (Pham, Sheu ve Nagai, 2015; Sheu ve diğerleri., 2014a).

Öğrenci-Madde Çizelgesi Analizi sonucunda iki tür indeks üretilir. Bunlar, öğrenci uyarı indeksi (Student Caution Indices) ve madde uyarı indeksi (Problem Caution Indices) olarak adlandırılmaktadır.

Madde Uyarı İndeksi

Madde uyarı indeksi (MUİ) maddelerin belli kategoriler altında sınıflandırılmasını ve bu sayede her bir maddenin işlerliğiyle ilgili dönüt elde edilmesini sağlar. MUİ indeksi aşağıdaki formülle hesaplanmaktadır (Lin ve Yih, 2015; Sheu ve diğerleri., 2014b).

$$MU\dot{I} = 1 - \frac{\sum_{n=1}^{N}(Ynm)(Yn) - (Ym).(\mu')}{\sum_{n=1}^{Ym} Yn - (Ym).(\mu')}$$

m: Madde değişkeni (m=1,2,3,…M)

n: Öğrenci değişkeni (n=1,2,3,…N)

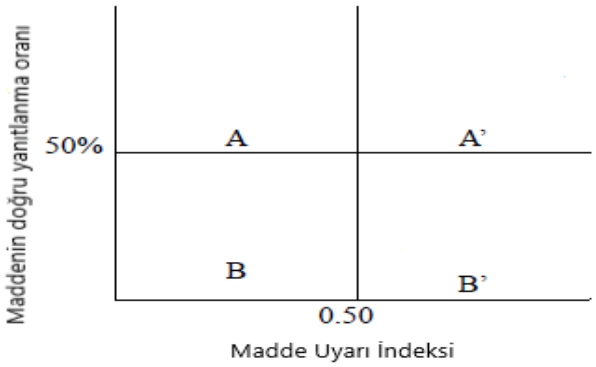
Ynm: n. öğrencinin m. maddeden aldığı puan (doğru:1, yanlış: 0)

Ym: Öğrencilerin m. maddeden aldıkları puanların toplamı

Yn: n. öğrencinin toplam test puanı

$\mu' : \frac{\sum_{n=1}^{N} Yn}{N}$

Madde uyarı indeksi (MUİ), maddeleri dört farklı sınıfa atamak için kullanılmaktadır. İndeks 0 ile 1 aralığında değer almaktadır. Herhangi bir madde için hesaplanan indeks değerinin 0,50'den daha büyük bir değer alması, ilgili maddenin işlerliğiyle ilgili bazı sorunların olduğu yönünde yorumlanmaktadır. MUİ değeri maddenin doğru yanıtlanma oranı ile birlikte yorumlanır. Bu iki ölçünün alacağı değerlere göre maddenin hangi sınıfa atanacağına ilişkin görsel Şekil 1'de sunulmuştur.



Şekil 1. MUİ ve oran değerlerine göre maddelerin atanacağı sınıflar.

Şekil 1'de görüldüğü üzere MUİ değerinin 0,50'den düşük bir değer aldığı durumda maddenin doğru yanıtlanma oranı %50'den yüksek ise A; düşük ise B sınıfına atandığı ve MUİ değerinin 0,50'den büyük bir değer aldığı durumda maddenin doğru yanıtlanma oranı %50'den yüksekse A'; düşük ise B' sınıfına atandığı belirtilebilir. Yapılan bu sınıflandırmalara ilişkin yapılabilecek yorumlar aşağıda verilmiştir (Sheu, Pham, Nguyen ve Nguyen, 2013).

A: Düzgün çalışan bir madde.

A': Kısmi olarak sorunlu bir madde. Tekrar gözden geçirilebilir.

B: Zor bir madde.

B':Sorunlu bir madde. Ters işliyor olabilir. Testten atılmalı ya da yeniden yazılmalı.

Öğrenci Uyarı İndeksi

Öğrenci uyarı indeksi (ÖUİ) öğrencilerin belli kategoriler altında sınıflandırılmasını ve bu sayede her bir öğrencinin testle yoklananbilgi/becerileri öğrenme ve maddeleri dikkatli bir şekilde yanıtlama durumları hakkında dönüt elde edilmesini sağlar. ÖUİ indeksi aşağıdaki formülle hesaplanmaktadır (Lin ve Yih, 2015; Sheu ve diğerleri., 2014a).

$$\text{ÖUİ} = 1 - \frac{\sum_{m=1}^{M}(Ynm)(Ym) - (Yn).(u')}{\sum_{m=1}^{Yn} Ym - (Yn).(u')}$$

m: Madde değişkeni (m=1,2,3,…M)

n: Öğrenci değişkeni (n=1,2,3,…N)

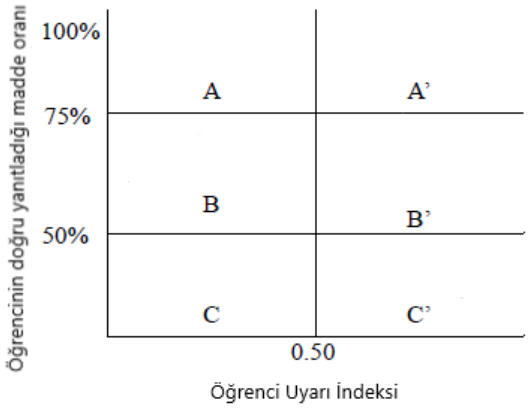
Ynm: n. öğrencinin m. maddeden aldığı puan (doğru:1, yanlış: 0)

Ym: Öğrencilerin m. maddeden aldıkları puanların toplamı

Yn: n. öğrencinin toplam test puanı

u': $\frac{\sum_{m=1}^{M} Ym}{M}$

Öğrenci uyarı indeksi (ÖUİ), öğrencileri altı farklı sınıfa atamak için kullanılmaktadır. İndeks 0 ile 1 aralığında değer almaktadır. Herhangibir öğrenci için hesaplanan indeks değerinin 0,50'den daha büyük bir değer alması, o öğrencinin öğrenme eksikliklerinin olabileceği ya da bazı maddeleri doğru cevaplayabilecekken dikkatsizlikle kaçırmış olabileceği uyarısını vermektedir. ÖUİ değeri öğrencinin doğru yanıt verdiği madde sayısının testte yer alan madde sayısına oranı ile birlikte yorumlanır. Bu iki ölçünün alacağı değerlere göre öğrencinin(bireyin) hangi sınıfa atanacağına ilişkin görsel Şekil 2'de sunulmuştur.



Şekil 2. ÖUİ ve oran değerlerine göre öğrencilerin atanacağı sınıflar.

Şekil 2'de yer alan değerler doğrultusunda yapılan sınıflandırmalar ve bunlara ilişkin yapılabilecek yorumlar aşağıda verilmiştir (Sheu ve diğerleri., 2013).

A: Etkili öğrenme gerçekleşmiş.

A': Öğrenme var fakat çok dikkatsiz.

B: Genel olarak iyi ancak biraz daha fazla çalışmaya ihtiyacı var.

B': Biraz dikkatsiz ve daha fazla çalışmaya ihtiyacı var.

C: Öğrenme düzeyi yetersiz.

C': Öğrenme gerçekleşmemiş.

Yapılan bu sınıflandırmaların Sato Test Kuramı'nın biçimlendirici değerlendirme ve öğretimi geliştirme amacıyla kullanılmasında öğretmenlere önemli bir referans oluşturduğu ifade edilebilir. Bu sınıflandırma sayesinde öğretmenler, farklı kesme puanları belirlemek gibi bir zorlukla karşı karşıya kalmaksızın öğrencileri altı farklı gruba ayırabilmekte ve öğretimi bu grupların ihtiyaçlarına uygun olarak düzenleyebilmektedirler.

ÖUİ ve MUİ indeksleriyle ulaşılan nitel veri düzeyindeki sonuçlar, yetenek düzeyleri bakımından bireyleri ve güçlükleri bakımından maddeleri karşılaştırabilme noktasında sınırlı kalmaktadır. Ö-MÇA'nın bu sınırlılığını ortadan kaldırmak amacıyla geliştirilen Gri Ö-MÇA'ya dayalı olarak öğrencilerin test performans düzeylerini belirlemek amacıyla "öğrenciler için sınırlandırılmış gri ilişki derecesi" –ÖİSGİD- (Localized Grey Relational Grade-Student) ve madde güçlüklerini belirlemek amacıyla "maddeler için sınırlandırılmış gri ilişki derecesi" -MİSGİD- (Localized Grey Relational Grade-Problem) hesaplanmaktadır (Sheu ve diğerleri., 2014a).

Öğrenciler için Sınırlandırılmış Gri İlişki Derecesi

Öğrenciler için Sınırlandırılmış Gri İlişki Derecesi (ÖİSGİD) değerleri öğrencilerin yetenek ya da başarı düzeylerine ilişkin bilgi vermektedir. ÖİSGİD hesaplanırken kullanılan vektörler ve formüller aşağıda verilmiştir (Sheu ve diğerleri., 2014c).

Gri ilişki analizinde karşılaştırılan vektörler:

$x_0 = (x_0(1), x_0(2), \dots, x_0(k), \dots, x_0(n))$: Referans vektör

$x_i = (x_i(1), x_i(2), \dots, x_i(k), \dots, x_i(n))$: İncelenen vektör

Gri öğrenci-madde çizelgesi analizinde ÖİSGİD değeri elde etmek için incelenen vektör:

$S_i = (x_i(1), x_i(2), \dots, x_i(k), \dots, x_i(n))$: i. öğrencinin tüm maddelere verdiği yanıtlar vektörü, k = 1,2,3,…,n

$$\text{ÖİSGİD} = \Gamma_{(x_0, x_i)} = \frac{\overline{\Delta}_{max} - \overline{\Delta}_{0i}}{\overline{\Delta}_{max} - \overline{\Delta}_{min}}$$

$\overline{\Delta}_{0i} = (\sum_{k=1}^{n}(x_0(k) - x_i(k)^p)^{\frac{1}{p}}$ *Referans vektör ve incelenen vektör (öğrenciler için) arasındaki Minkowski uzaklık ölçüsü.*

$\overline{\Delta}_{max} = \overline{\Delta}_{0i}$ *için maksimum değer.*

$\overline{\Delta}_{min} = \overline{\Delta}_{0i}$ *için minimum değer.*

ÖİSGİD değerlerine dayalı olarak öğrencilerin/bireylerin ölçülen özellikteki performansları aşağıdaki gibi değerlendirilebilir.

Tablo 1. *ÖİSGİD değerlendirme ölçütleri*

Bireyin Performansı	ÖİSGİD değeri*
Yüksek	ÖİSGİD ≥ 0,75
Ortanın üstü	0,75 > ÖİSGİD ≥ 0,50
Orta	0,50 > ÖİSGİD ≥ 0,25
Düşük	0,25 > ÖİSGİD

*ÖİSGİD 0 ile 1 aralığında değer almaktadır.

ÖİSGİD değerleri, öğrencilerin düzeylerini belirlemeye ve buna bağlı olarak onlar hakkında kararlar almaya olanak tanıyan, karşılaştırılabilir ve sıralanabilir nicel verilerdir.

Maddeler için Sınırlandırılmış Gri İlişki Derecesi

Maddeler için sınırlandırılmış gri ilişki derecesi (MİSGİD) maddelerin güçlük düzeylerine ilişkin bilgi vermektedir. MİSGİD 0 ile 1 aralığında değer almakta ve kolay maddeler için yüksek, zor maddeler için düşük değerler elde edilmektedir. Başka bir ifadeyle, MİSGİD değeri ne denli yüksekse

madde o denli kolaydır, denilebilir. MİSGİD hesaplanırken kullanılan vektörler ve formüller aşağıda verilmiştir (Sheu ve diğerleri., 2014c).

Gri öğrenci-madde çizelgesi analizinde MİSGİD değeri elde etmek için incelenen vektör:

$P_k = (x_1(k), x_2(k), \ldots, x_i(k), \ldots, x_m(k))$: k.maddeye tüm öğrencilerin verdiği yanıtlar vektörü, i= 1,2,3,…,m

$$MİSGİD = \Gamma_{(x_j, x_{0j})} = \frac{\overline{\Delta}_{max} - \overline{\Delta}_{k0}}{\overline{\Delta}_{max} - \overline{\Delta}_{min}}$$

$\overline{\Delta}_{k0} = (\sum_{i=1}^{m}(x_0(k) - x_i(k)^P)^{\frac{1}{P}}$ *Referans vektör ve incelenen vektör (maddeler için) arasındaki Minkowski uzaklık ölçüsü.*

$\overline{\Delta}_{max} = \overline{\Delta}_{k0}$ *için maksimum değer.*

$\overline{\Delta}_{min} = \overline{\Delta}_{k0}$ *için minimum değer.*

STK ile ilgili çalışmaların son yıllarda özellikle Uzak Doğu Coğrafyasında bir artış gösterdiği ve yapılan İngilizce yayınların, kuramı tanıtma (Lin ve Chen, 2006; Wang, Sheu ve Nagai, 2011; Wang ve Chen 2013), kuramın uygulamasına yönelik geliştirilen bir yazılımı tanıtma (Pham, Sheu ve Nagai, 2015; Sheu ve diğerleri., 2013; Sheu ve diğerleri., 2014b; Sheu ve diğerleri., 2014a; Wu, 1999), STK aracılığıyla kavram yanılgılarına neden olan maddeleri ve/veya kavram yanılgılarına sahip olan öğrencileri saptama (Sheu ve diğerleri., 2013; Tsai ve diğerleri., 2014) ve STK ve MTK'nın karşılaştırılması (Sheu ve diğerleri., 2014b; Tatsuoka, 1984) alanlarıyla ilişkili olduğu belirlenmiştir. Bu belirlemeden hareketle, Gri İlişki Analizi'nin kurama eklendiği 2010 yılından sonra dahi Avrupa ve Amerika'da kurama olan ilginin zayıf kaldığına ilişkin bir çıkarım yapılabilir. Bu anlamda, farklı katılımcı grupları üzerinde STK çerçevesinde üretilen madde ve birey özelliklerine ilişkin değerlerin diğer test kuramları aracılığıyla elde edilen değerlerle karşılaştırılmasının hem farklı kuramlar arasındaki benzerlik ve farklılıkları ortaya koyması bakımından hem de STK'nın zayıf ve güçlü yönlerini ortaya çıkarması bakımından çalışmanın ilgili alanyazına katkı sunacağı düşünülmektedir.

Yöntem

Bu bölümde, araştırmanın modeli, çalışma grubu, veri seti ve verilerin analizine ilişkin bilgiler verilmiştir.

Araştırma Modeli

Bu araştırmada, Öğrenci Başarılarının Belirlenmesi Sınavı'na (ÖBBS-2005) katılan öğrencilerin matematik alt testine verdikleri yanıtlar üzerinden, madde analizlerinin Sato Test Kuramı (STK), Klasik Test Kuramı (KTK) ve Madde Tepki Kuramı (MTK) ile yapılarak elde edilen madde özellikleri arasındaki ilişki düzeylerinin, testlerden çıkarılmasına ya da düzeltilmesine karar verilen sorunlu

maddelerin belirlenmesi ve sözü edilen kuramlara dayalı olarak elde edilen öğrenci başarı puanlarının karşılaştırılması amaçlanmıştır. Bu temel amaçtan hareketle aşağıdaki sorulara yanıt aranmıştır.

ÖBBS (2005) matematik alt testinde yer alan maddeler için random belirlenen 50'şer kişilik 10, 100'er kişilik 10, 200'er kişilik 10 ve 600'er kişilik 10 farklı örneklem üzerinden hesaplanan (Madde Tepki Kuramı'na dayalı kestirimler için bu örneklemler değil tam veri seti kullanılmıştır),

1. Sato Test Kuramı'nın madde uyarı indeksleri (MUİ) ile Klasik Test Kuramı'nın madde ayırıcılık indeksleri (Çift Serili Korelasyon Katsayıları) arasındaki korelasyonlar ne düzeydedir?
2. Sato Test Kuramı'nın maddeler için sınırlandırılmış gri ilişki dereceleri (MİSGİD) ile Klasik Test Kuramı'nın madde güçlük indeksleri arasındaki korelasyonlar ne düzeydedir?
3. Sato Test Kuramı'nın öğrenciler için sınırlandırılmış gri ilişki dereceleri (ÖİSGİD) ile Klasik Test Kuramı'nın öğrenci başarı puanları (toplam test puanları) arasındaki korelasyonlar ne düzeydedir?
4. Sato Test Kuramı'nın madde uyarı indeksleri (MUİ) ile Madde Tepki Kuramı'nın *a* parametreleri arasındaki korelasyonlar ne düzeydedir?
5. Sato Test Kuramı'nın maddeler için sınırlandırılmış gri ilişki dereceleri (MİSGİD) ile Madde Tepki Kuramı'nın *b* parametreleri arasındaki korelasyonlar ne düzeydedir?
6. Sato Test Kuramı'nın öğrenciler için sınırlandırılmış gri ilişki dereceleri (ÖİSGİD) ile Madde Tepki Kuramı'nın yetenek parametreleri arasındaki korelasyonlar ne düzeydedir?
7. Klasik Test Kuramı, Madde Tepki Kuramı ve Sato Test Kuramı'na dayalı olarak yürütülen madde analizlerinin sonuçlarından hareketle testten çıkarılması ya da düzeltilmesi gerektiğine karar verilen maddeler ne oranda uyum göstermektedir?

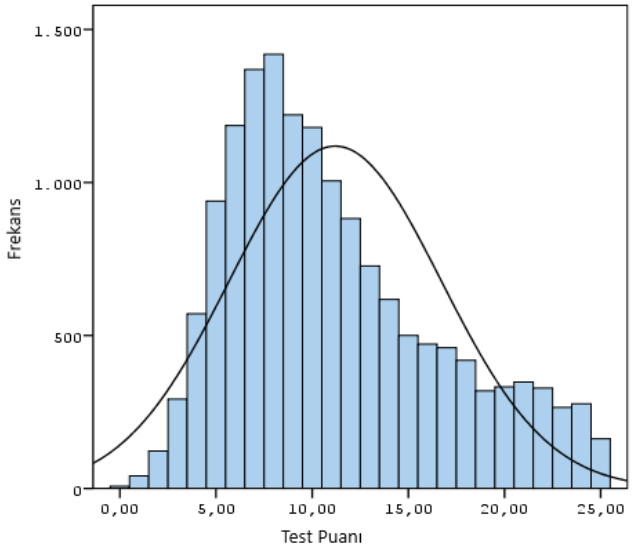
Araştırmada ele alınan yöntemlerin araştırma verisi üzerinden işleyişlerinin incelenmesi amaçlandığından kullanılan verinin güncel olup olmaması önem taşımamaktadır. Araştırma, analiz sonuçlarının karşılaştırılması yoluyla yöntemler arasındaki bağlantıların belirlenmesi ve buradan hareketle genel bir tablonun ortaya koyulması bakımından betimsel bir araştırma niteliği taşımaktadır.

Çalışma Grubu ve Veri Seti

Çalışma grubunda, 2005 yılında uygulanan Öğrenci Başarılarının Belirlenmesi Sınavı'na (ÖBBS) katılan 8.sınıf öğrencilerinden oluşan 15461 kişi yer almaktadır. Araştırma verileri, ÖBBS matematik alt testinde yer alan çoktan seçmeli maddelere verilen yanıtların ikili puanlanması (1-0) ile oluşturulmuştur.

ÖBBS'nin bu araştırma kapsamında ele alınan matematik alt testinde dört seçenekli 25 madde yer almaktadır. Sınava katılan öğrencilerin matematik testinden aldıkları puanların (doğru yanıtlanan

maddeler için 1 yanlış yanıtlanan ve boş bırakılan maddeler için 0 puan verilmiştir) histogram grafiği Şekil 3'te sunulmuştur.



*Şekil 3.*Matematik test puanlarının dağılımına ilişkin histogram grafiği.

Şekil 3'te verilen test puanları dağılımının ortalaması 11,199 ve ortancası 10,000 olarak hesaplanmıştır. Dağılımın standart sapması 5,513'tür. Dağılıma ilişkin çarpıklık katsayısı 0,697'dir ve test puanları bu anlamda sağa çarpık bir dağılım özelliği göstermektedir. Basıklık katsayısı ise -0,320 olarak hesaplanmıştır.

Verilerin Analizi

Bu bölümde öncelikle, araştırma verisinin yapılacak olan analizlerin gerektirdiği varsayımları karşılayıp karşılamadığı incelenmiş daha sonra veri çözümleme süreci, araştırma sorularını sırasıyla ele alacak şekilde özetlenmeye çalışılmıştır.

Varsayımların Test Edilmesi

Bu aşamada ilk olarak hem KTK hem de MTK için gerekli olan tek boyutluluk varsayımı Paralel Analiz yöntemi ile incelenmiştir. Söz konusu yöntem pek çok araştırmacı tarafından faktör (boyut) sayısının belirlenmesinde en iyi yöntem olarak nitelendirilmektedir (Fabrigar, Wegener, MacCallum ve Strahan, 1999; Field, 2009; Hayton, Allen ve Scarpello, 2004; Henson ve Roberts, 2006; Horn, 1965; Thompson, 2004; Zwick ve Velicer, 1986). Bu amaçla, "Factor 10" programı aracılığıyla tetrakorik korelasyon matrisine dayalı faktör analizi yürütülmüş ve faktörleştirme yöntemi olarak Coughlin (2013) tarafından önerilen Ağırlıklandırılmamış En Küçük Kareler Yöntemi (Unweighted Least Squares) seçilmiştir. Korelasyon matrisinin faktörleşebilirliğine ilişkin kanıtlar Bartlett Küresellik Testi (Bartlett's Test of Sphericity) ve Kaiser-Meyer-Olkin Örneklem Uygunluğu Ölçütü (Kaiser-Meyer-Olkin Test of Sampling Adeqacy) ile elde edilmiştir. Matematik alt testi için Bartlett

Küresellik Testi sonuçları korelasyon matrisi ile birim matris arasında istatistiksel olarak manidar bir farklılık olduğuna (p= 0.000) işaret etmektedir. Teste ilişkin KMO değeri (0,901) göz önünde bulundurulduğunda, korelasyon matrislerinin faktörleşebilirlik düzeylerinin "mükemmel" olduğu şeklinde bir değerlendirmede bulunulabilir (Beavers ve diğerleri., 2013). Korelasyon matrisine ilişkin varsayımlar sağlandıktan sonra yapılan Faktör Analizi ve buna dayalı olarak yapılan Paralel Analiz sonuçları, test için "tek boyut" önermiştir. Elde edilen bulgulara dayalı olarak araştırma verisinin tek boyutluluk varsayımını karşıladığına karar verilmiştir.

Tek boyutluluk, bir diğer MTK varsayımı olan yerel bağımsızlığın varlığına da kanıt olarak gösterilebilir. Bu kanıtın yanı sıra IRTPRO programı aracılığıyla maddeler arası standardize edilmiş ki-kare değerleri incelenmiş ve az sayıda madde arasında yüksek ki-kare değerlerinin tespit edilmesine karşın testin bütünsel olarak ele alındığı durumda maddeler arası ilişkilerin yerel bağımsızlığı ihlal edecek düzeyde olmadığı sonucuna ulaşılmıştır.

Araştırma verisine ilişkin temel varsayımların kontrolünden sonra MTK parametrelerinin daha doğru kestirilebilmesi için veriye en iyi uyum sağlayan MTK modelinin seçilmesi aşamasına geçilmiştir. Bu amaçla, teste ait veriler 1P, 2P ve 3P Lojistik Modellere göre analiz edilmiş ve elde edilen -2log Likelihood değerleri arasındaki farklar incelenmiştir. Söz konusu değerler Tablo 2'de sunulmuştur.

Tablo 2. *Lojistik modeller için -2log likelihood değerleri*

Test	*1PLM*	*2PLM*	*3PLM*
Matematik	471051,75	468898,42	463975,82

Modellerin uyum düzeylerinin karşılaştırılmasında χ^2 fark testi kullanılmıştır. Test için en uygun modelin Üç Parametreli Lojistik Model olduğu sonucuna ulaşılmıştır. Yapılan hesaplamalar aşağıdaki şekilde ifade edilebilir.

$\chi^2 = -2\log Likelihood_{1plm} - (-2\log Likelihood_{2plm}) = 471051{,}75 - 468898{,}42 = 2153{,}33 > 37{,}65\ (\chi^2_{(df=25\ p=;0{,}05)})$ ve,

$\chi^2 = -2\log Likelihood_{2plm} - (-2\log Likelihood_{3plm}) = 468898{,}42 - 463975{,}82 = 4922{,}60 > 37{,}65\ (\chi^2_{(df=25\ p=;0{,}05)})$

Seçilen modelin veriye uygunluğuna ilişkin daha fazla kanıt elde edebilmek amacıyla 3PLM ile üretilen madde ve yetenek parametrelerinin değişmezliğine yönelik incelemeler yapılmıştır. Madde parametrelerinin değişmezliğine ilişkin kanıt elde edilmesi amacıyla matematik testini yanıtlayan bireylere ait yanıt örüntüleri seçkisiz olarak iki gruba ayrılmış ve her iki gruptan kestirilen madde parametreleri arasındaki ilişki Spearman-Brown Sıra Farkları Korelasyon Katsayıları ile incelenmiştir. İki gruptan elde edilen madde parametreleri arasındaki korelasyonlar a parametreleri

için 0,926; b parametreleri için 0,993; c parametreleri için 0,940 olarak belirlenmiştir. Bu bulguya dayalı olarak madde parametrelerinin 3PLM'de değişmezlik özelliği gösterdiği yorumu yapılabilir.

Yetenek parametrelerinin değişmezliğinin incelenmesi amacıyla, matematik testinde yer alan maddeler rastgele bir şekilde ikişer gruba ayrılarak alt testler oluşturulmuş ve bu alt testlerden 3PLM ile yetenek parametreleri kestirilmiştir. Test maddelerinden oluşturulan iki ayrı alt test üzerinden kestirilen yetenek parametreleri arasındaki korelasyon düzeyi Pearson Momentler Çarpımı Korelasyon Katsayısı tekniği ile incelenmiş ve bu değer 0,887 olarak belirlenmiştir. Söz konusu bulguya dayalı olarak, testi yanıtlayan bireylerin yetenek düzeylerinin 3PLM ile madde örnekleminden bağımsız olarak kestirilebildiği iddiasında bulunulabilir. Yapılan incelemeler sonrasında, araştırma verisinin çözümlemeler için gerekli olan varsayımları karşıladığına karar verilmiştir. Bununla birlikte, MTK'ya dayalı çözümlemelerde araştırma verisine en iyi uyum sağlayan Üç Parametreli Lojistik Modelin kullanılması ve diğer test kuramlarıyla yapılan karşılaştırmaların bu modele dayalı olarak kestirilen parametrelerle sınırlandırılması kararlaştırılmıştır.

Veri Çözümleme Süreci

Öncelikle, ÖBBS'ye (2005) katılan 15461 8.sınıf öğrencisinin matematik testinden elde edilen yanıt örüntüleri arasından basit seçkisiz örnekleme yöntemiyle her birinden 10'ar adet olmak üzere 50'şer, 100'er, 200'er ve 600'er kişilik örneklemler oluşturularak araştırma verisi araştırma problemini çözümlemeye uygun hale getirilmiştir.

Araştırmanın birinci sorusuna yanıt aramak amacıyla, oluşturulan tüm örneklemler üzerinden test maddelerinin r_{jx} ve MUİ değerleri; ikinci sorusuna yanıt aramak amacıyla p ve MİSGİD değerleri elde edilmiştir. Söz konusu değerler arasındaki korelasyon katsayıları her bir örneklem büyüklüğü için 10 farklı örneklem üzerinden Spearman'ın Sıra Farkları Korelasyon Katsayısı Tekniği ile hesaplanarak ortancaları alınmıştır.

Araştırmanın üçüncü sorusuna yanıt aramak amacıyla, oluşturulan örneklemler içerisinde yer alan öğrencilerin test toplam puanları ve öğrenciler için sınırlandırılmış gri ilişki dereceleri (ÖİSGİD) elde edilmiştir. Söz konusu değerler arasındaki korelasyon katsayıları her bir örneklem büyüklüğü için 10 farklı örneklem üzerinden Pearson Korelasyon Katsayısı Tekniği ile hesaplanarak ortancaları alınmıştır.

Araştırmanın dördüncü ve beşinci sorularına yanıt aramak amacıyla, testte yer alan maddeler için 50, 100, 200 ve 600'er kişilik 10'ar farklı örneklem üzerinden MUİ ve MİSGİD değerleri ile teste katılan tüm öğrenciler üzerinden 3PLM ile a ve b parametreleri belirlenmiş, ilgili değişkenler arasındaki korelasyon katsayıları Spearman'ın Sıra Farkları Korelasyon Katsayısı Tekniği ile hesaplanarak ortancaları alınmıştır.

Araştırmanın altıncı sorusuna yanıt aramak amacıyla, testi yanıtlayan ve 50, 100, 200 ve 600'er kişilik örneklemler içerisinde yer alan öğrencilerin ÖİSGİD değerleri ile aynı öğrencilerin 3PLM ile kestirilen yetenek (θ) parametreleri (θ kestirimleri EAP yöntemi ile tüm veri seti üzerinden gerçekleştirilmiştir.) arasındaki korelasyonlar Pearson Korelasyon Katsayısı Tekniği ile incelenmiştir.

Araştırmanın yedinci sorusuna yanıt aramak amacıyla, oluşturulan örneklemler üzerinden Çift Serili Korelasyon Katsayısı tekniği ile madde ayırıcılık indeksleri (r_{jx}) elde edilerek sorunlu maddeler KTK bağlamında belirlenmiştir. Daha sonra, aynı örneklemlerden alınan veriler Sato Test Kuramı'na dayalı olarak analiz edilmiş ve sorunlu maddeler MUİ'ler aracılığı ile yapılan sınıflandırmalar yoluyla belirlenmiştir. Son olarak, Madde Tepki Kuramı'nın *a, c* parametreleri ve madde bilgi fonksiyonları aracılığı ile sorunlu maddeler belirlenerek her üç kurama göre ulaşılan sonuçlar arasında karşılaştırmalar yapılmıştır.

Verilerin analizi süresince Sato Test Kuramına dayalı kestirimler Sheu, Pham, Nguyen ve Nguyen (2013) tarafından geliştirilen MATLAB eklentisi ile Madde Tepki Kuramına dayalı parametreler IRTPRO (V, 4.2) yazılımı ile Klasik Test Kuramına dayalı istatistikler ise TAP (V,14.7.4) yazılımı ile elde edilmiştir. Spearman'ın Sıra Farkları Korelasyon Katsayısı Tekniği, Pearson Korelasyon Katsayısı Tekniği ve Kruskal-Wallis-H Testi ile yapılacak olan karşılaştırmalar için ise SPSS (V,22) yazılımı kullanılmıştır.

Araştırmanın Etik İzinleri

Bu çalışmada "Yükseköğretim Kurumları Bilimsel Araştırma ve Yayın Etiği Yönergesi" kapsamında uyulması belirtilen tüm kurallara uyulmuş, yönergenin ikinci bölümünde yer alan "Bilimsel Araştırma ve Yayın Etiğine Aykırı Eylemler" başlığı altında belirtilen eylemlerden hiçbiri gerçekleştirilmemiştir.

Etik kurul izin bilgileri: Hacettepe Üniversitesi Eğitim Bilimleri Enstitüsü'nün 2019 yılında doktora tezi olarak kabul ettiği araştırmanın bir bölümünü içeren bu makalede yapılan işlemler için çalışmada hazır veri kullanıldığı gerekçesiyle danışman onayıyla aynı enstitüye etik komisyon muafiyeti başvurusu yapılmış ve ilgili araştırma etik kurul izninden muaf tutulmuştur.

Muafiyet İzin Formu Tarihi= 23/06/2016

Bulgular ve Yorumlar

"Klasik Test Kuramı'nın madde ayırıcılık indeksleri (Çift Serili Korelasyon Katsayıları) ile Sato Test Kuramı'nın madde uyarı indeksleri (MUİ) arasında ne düzeyde korelasyon bulunmaktadır?" sorusuna yanıt bulmak amacıyla ÖBBS matematik alt testi maddeleri için 50, 100, 200 ve 600'er kişilik örneklemler üzerinden r_{jx} ve MUİ değerleri elde edilmiş ve her bir örneklem büyüklüğü için 10 farklı örneklem üzerinden elde edilen değerler arasındaki korelasyon katsayıları hesaplanarak ortancaları alınmıştır. Konuya ilişkin bulgular Tablo 3'te sunulmuştur.

Tablo 3. *Matematik testi madde ayırıcılık indeksleri ile madde uyarı indeksleri arasındaki korelasyonlara ilişkin bulgular*

Örneklem Büyüklüğü	50	100	200	600
Korelasyon$_{(MUİ-r_{jx})}$	-0,968*	-0,934*	-0,944*	-0,934*

*Sunulan değerler her bir örneklem büyüklüğü için 10 farklı örneklem üzerinden elden edilen korelasyon katsayılarının ortancaları alınarak elde edilmiştir.

Tablo 3 incelendiğinde, ele alınan tüm örneklem büyüklükleri için KTK'nın madde ayırıcılık indeksleri ile STK'nın madde uyarı indeksleri arasında negatif yönde ve yüksek düzeyde korelasyonların olduğu görülmektedir. Elde edilen yüksek korelasyon katsayılarından hareketle, matematik testinde yer alan maddelerin ayırt edicilik düzeylerinin belirlenebilmesi bakımından, madde ayırıcılık ve madde uyarı indekslerinin benzer şekilde çalıştığı görüşü öne sürülebilir. Bununla birlikte, maddelerin r_{jx} ve MUİ değerleri arasındaki ilişki düzeylerinin örneklem büyüklüğünden etkilendiğine dair bir bulguya ulaşılmamıştır.

"Klasik Test Kuramı'nın madde güçlük indeksleri ile Sato Test Kuramı'nın maddeler için sınırlandırılmış gri ilişki dereceleri (MİSGİD) arasında ne düzeyde korelasyon bulunmaktadır?" sorusuna yanıt bulmak amacıyla ÖBBS matematik alt testi maddeleri için 50, 100, 200 ve 600'er kişilik örneklemler üzerinden p ve MİSGİD değerleri elde edilmiştir. Söz konusu değerler arasındaki korelasyon katsayıları her bir örneklem büyüklüğü için 10 farklı örneklem üzerinden hesaplanarak ortancaları alınmıştır. Konuya ilişkin bulgular Tablo 4'te sunulmuştur.

Tablo 4. *Matematik testi madde güçlük indeksleri ile maddeler için sınırlandırılmış gri ilişki dereceleri arasındaki korelasyonlara ilişkin bulgular*

Örneklem Büyüklüğü	50	100	200	600
Korelasyon$_{(MİSGİD-p)}$	1,000*	0,998*	0,998*	0,999*

*Sunulan değerler her bir örneklem büyüklüğü için 10 farklı örneklem üzerinden elden edilen korelasyon katsayılarının ortancaları alınarak elde edilmiştir.

Tablo 4 incelendiğinde, ele alınan tüm örneklem büyüklükleri için KTK'nın madde güçlük indeksleri ile STK'nın maddeler için sınırlandırılmış gri ilişki dereceleri arasında pozitif yönde ve yüksek düzeyde korelasyonların olduğu görülmektedir. Söz konusu korelasyonların örneklem büyüklüğünden etkilendiğine dair bir bulguya ulaşılmamıştır. Elde edilen yüksek korelasyon katsayılarından hareketle, matematik testinde yer alan maddelerin güçlük düzeylerinin belirlenebilmesi bakımından, madde güçlük indekslerinin ve maddeler için sınırlandırılmış gri ilişki derecelerinin benzer şekilde çalıştığı görüşü öne sürülebilir.

"Klasik Test Kuramı'nın öğrenci başarı puanları (toplam test puanları) ile Sato Test Kuramı'nın öğrenciler için sınırlandırılmış gri ilişki dereceleri (ÖİSGİD) arasında ne düzeyde korelasyon bulunmaktadır?" sorusuna yanıt bulmak amacıyla ÖBBS matematik alt testini yanıtlayan ve 50, 100, 200 ve 600'er kişilik örneklemler içerisinde yer alan öğrencilerin test toplam puanları ve öğrenciler için sınırlandırılmış gri ilişki dereceleri elde edilmiştir. Söz konusu değerler arasındaki

korelasyon katsayıları her bir örneklem büyüklüğü için 10 farklı örneklem üzerinden hesaplanarak ortancaları alınmıştır. Konuya ilişkin bulgular Tablo 5'de sunulmuştur.

Tablo 5. *Matematik testi toplam puanları ile öğrenciler için sınırlandırılmış gri ilişki dereceleri arasındaki korelasyonlara ilişkin bulgular*

Örneklem Büyüklüğü	50	100	200	600
Korelasyon$_{(\text{ÖİSGİD}-\text{TTP})}$	0,979*	0,977*	0,978*	0,976*

*Sunulan değerler her bir örneklem büyüklüğü için 10 farklı örneklem üzerinden elden edilen korelasyon katsayılarının ortancaları alınarak elde edilmiştir.

Tablo 5 incelendiğinde, ele alınan tüm örneklem büyüklükleri için test toplam puanları ile öğrenciler için sınırlandırılmış gri ilişki dereceleri arasında pozitif yönde ve yüksek düzeyde korelasyonların olduğu görülmektedir. Söz konusu korelasyonların örneklem büyüklüğünden etkilendiğine dair bir bulguya ulaşılmamıştır. Elde edilen yüksek korelasyon katsayılarından hareketle, bireylerin matematik testinde gösterdikleri performansların belirlenmesi noktasında ÖİSGİD ve test toplam puanlarının benzerlik gösterdiği ifade edilebilir. Araştırmanın birinci, ikinci ve üçüncü alt problemine ilişkin bulgular değerlendirildiğinde STK ve KTK'nın psikometrik özellikler bakımından birbirine yakın sonuçlar ortaya koyduğu görülmüştür.

"Sato Test Kuramı'nın madde uyarı indeksleri (MUİ) ile Madde Tepki Kuramı'nın *a* parametreleri arasında ne düzeyde ilişki bulunmaktadır?" Sorusuna yanıt bulmak amacıyla ÖBBS matematik alt testi maddeleri için 50, 100, 200 ve 600'er kişilik 10'ar farklı örneklem üzerinden elde edilen MUİ değerleri ile teste katılan tüm öğrenciler üzerinden 3PLM ile elde edilen a parametreleri arasındaki korelasyon katsayıları incelenmiştir. Elde edilen bulgular Tablo 6'da sunulmuştur.

Tablo 6. *Matematik testi madde uyarı indeksleri ile a parametreleri arasındaki korelasyonlara ilişkin bulgular*

Örneklem Büyüklüğü	50	100	200	600
Korelasyon$_{(\text{MUİ}-\text{a})}$	-0,082*	-0,154*	-0,138*	-0,269*

*Sunulan değerler her bir örneklem büyüklüğü için 10 farklı örneklem üzerinden elden edilen korelasyon katsayılarının ortancaları alınarak elde edilmiştir.

Tablo 6 incelendiğinde, matematik testi maddeleri için elde edilen STK'nın madde uyarı indeksleri ve MTK'nın *a* parametreleri arasında tüm örneklem büyüklükleri için negatif yönde ve düşük düzeyde bir ilişkinin bulunduğu görülmektedir. Bu bakımdan, matematik testi maddelerinin ayırt ediciliklerinin belirlenmesi noktasında MUİ ve *a* parametrelerinin benzer şekilde çalışmadıkları ifade edilebilir.

"Sato Test Kuramı'nın maddeler için sınırlandırılmış gri ilişki dereceleri (MİSGİD) ile Madde Tepki Kuramı'nın *b* parametreleri arasında ne düzeyde ilişki bulunmaktadır?" Sorusuna yanıt bulmak amacıyla ÖBBS matematik alt testi maddeleri için 50, 100, 200 ve 600'er kişilik 10'ar farklı örneklem üzerinden elde edilen MİSGİD değerleri ile teste katılan tüm öğrenciler üzerinden 3PLM ile elde

edilen *b* parametreleri arasındaki korelasyon katsayıları incelenmiştir. Elde edilen bulgular Tablo 7'de sunulmuştur.

Tablo 1 *Matematik testi maddeler için sınırlandırılmış gri ilişki dereceleri ile b parametreleri arasındaki korelasyonlara ilişkin bulgular*

Örneklem Büyüklüğü	50	100	200	600
Korelasyon$_{(\text{MİSGİD}-b)}$	-0,805*	-0,880*	-0,891*	-0,928*

*Sunulan değerler her bir örneklem büyüklüğü için 10 farklı örneklem üzerinden elden edilen korelasyon katsayılarının ortancaları alınarak elde edilmiştir.

Tablo 7 incelendiğinde, STK'nın maddeler için sınırlandırılmış gri ilişki dereceleri ile MTK'nın *b* parametreleri arasında tüm örneklem büyüklükleri için negatif yönde ve yüksek düzeyde korelasyonların olduğu görülmektedir. Ayrıca, MİSGİD değerlerinin elde edildiği örneklem büyüklükleri arttıkça tüm katılımcıların yanıtları üzerinden elde edilen *b* parametreleri ile gösterdikleri korelasyon düzeylerinin de arttığı belirlenmiştir. Elde edilen bulgulardan hareketle, matematik testi maddelerinin güçlük düzeylerinin belirlenmesi noktasında MİSGİD ve *b* parametrelerinin benzerlik gösterdiği ifade edilebilir. Bu araştırmada, madde güçlüklerinin belirlenmesi noktasında STK ve MTK ile üretilen değerlerin yüksek düzeyde ilişki gösterdiğinin ortaya koyulmuş olması bakımından elde edilen bulguların Sheu ve diğerlerinin (2014c) yaptıkları araştırmada elde ettikleri bulgularla paralellik gösterdiği ifade edilebilir. Söz konusu araştırmada bu araştırmadan farklı olarak STK ile MTK modellerinden 1PLM karşılaştırılmıştır.

"Sato Test Kuramı'nın öğrenciler için sınırlandırılmış gri ilişki dereceleri (ÖİSGİD) ile Madde Tepki Kuramı'nın yetenek parametreleri arasında ne düzeyde korelasyon bulunmaktadır?" Soruna yanıt bulmak amacıyla, ÖBBS matematik alt testini yanıtlayan ve 50, 100, 200 ve 600'er kişilik örneklemler içerisinde yer alan öğrencilerin ÖİSGİD değerleri ile aynı öğrencilerin 3PLM ile kestirilen yetenek (θ) parametreleri (θ kestirimleri tüm veri seti üzerinden gerçekleştirilmiştir.) arasındaki korelasyon katsayıları incelenmiştir. Elde edilen bulgular Tablo 8'de sunulmuştur.

Tablo 8. *Matematik testini yanıtlayan öğrencilerin yetenek parametreleri ile öğrenciler için sınırlandırılmış gri ilişki dereceleri arasındaki korelasyonlara ilişkin bulgular*

Örneklem Büyüklüğü	50	100	200	600
Korelasyon$_{(\text{ÖİSGİD}-\boldsymbol{\theta})}$	0,936	0,915	0,941	0,949

Tablo 8 incelendiğinde, STK'nın öğrenciler için sınırlandırılmış gri ilişki dereceleri ile MTK'nın yetenek parametreleri arasında tüm örneklem büyüklükleri için pozitif yönde ve yüksek düzeyde korelasyonların olduğu görülmektedir. Söz konusu korelasyonların, ÖİSGİD değerlerinin elde edildiği örneklemlerin büyüklüğünden etkilendiğine dair net bir bulguya ulaşılamamıştır. Elde edilen yüksek korelasyon katsayılarından hareketle, bireylerin matematik testinde gösterdikleri performansların belirlenmesi noktasında ÖİSGİD ve θ değerlerinin benzerlik gösterdiği ifade edilebilir.

"Klasik Test Kuramı, Madde Tepki Kuramı ve Sato Test Kuramı'na dayalı olarak yürütülen madde analizlerinin sonuçlarından hareketle testten çıkarılması ya da düzeltilmesi gerektiğine karar verilen maddeler ne oranda uyum göstermektedir?" Sorusuna yanıt aramak amacıyla öncelikle matematik testi için 50, 100, 200 ve 600'er kişilik örneklemler üzerinden Çift Serili Korelasyon Katsayısı tekniği ile madde ayırıcılık indeksleri (r_{jx}) elde edilmiş ve elde edilen madde ayırıcılık indeksleri baz alınarak matematik testinden çıkarılması ya da düzeltilmesi gerektiğine karar verilen sorunlu maddeler ($r_{jx} < 0{,}30$) belirlenerek her bir örneklem büyüklüğü için Tablo 9'da sunulmuştur.

Tablo 9. *Matematik testindeki sorunlu maddelerin klasik test kuramı'na göre belirlenmesine ilişkin bulgular*

Örneklem Büyüklüğü	50	100	200	600
Sorunlu Madde Kümeleri	3, 5, 6, 11, 13, 14, 16, 22, 23	4, 6, 7, 16	10, 22, 23	4, 6, 7, 16, 23

Tablo 9 incelendiğinde, KTK'ya göre, farklı örneklem büyüklüklerinde en çok kez sorunlu olduğu tespit edilen maddelerin 6, 16 ve 23. maddeler olduğu söylenebilir.

Aynı örneklemlerden alınan veriler Sato Test Kuramı'na dayalı olarak analiz edilmiş ve maddeler MUİ'ler aracılığı ile sınıflandırılmıştır. Yapılan incelemeler sonucu matematik testinden çıkarılması ya da düzeltilmesi gerektiğine karar verilen sorunlu maddeler her bir örneklem büyüklüğü için Tablo 10'da sunulmuştur.

Tablo 10. *Matematik testindeki sorunlu maddelerin sato test kuramı'na göre belirlenmesine ilişkin bulgular*

Örneklem Büyüklüğü	50	100	200	600
Sorunlu Madde Kümeleri	3, 5, 6, 11, 13, 14, 16, 19, 22, 23	4, 6, 7, 10, 13, 16, 22, 23, 25	7, 10, 14, 22, 23	4, 7, 10,16, 22, 23

Tablo 10 incelendiğinde, 22. ve 23. maddelerin tüm örneklem büyüklüklerinde; 7. ve 10. maddelerin ise 100, 200 ve 600 kişilik örneklemlerde sorunlu olarak sınıflandırıldığı görülmektedir. STK ve KTK'ya dayalı olarak yapılan analizler sonrasında sorunlu olduğu tespit edilen madde kümeleri karşılaştırıldığında, örneklem büyüklüğü 50 iken 3, 5, 6, 11, 13, 14, 16, 22 ve 23. maddelerin; 100 iken 4, 6, 7 ve 16. maddelerin; 200 iken 10, 22 ve 23. maddelerin; 600 iken 4, 7, 16 ve 23. maddelerin ortak eleman olduğu belirlenmiştir. Yapılan karşılaştırmalardan hareketle, her iki kurama dayalı olarak sorunlu maddelerin tespiti amacıyla yürütülen analizlerin benzer sonuçlar ortaya koyduğu iddia edilebilir.

Madde Tepki Kuramı'na göre matematik testinden çıkarılması ya da düzeltilmesi gereken sorunlu maddeler a, *c* parametreleri ve madde bilgi fonksiyonları aracılığı ile belirlenmiştir. Bulgular, tüm katılımcılar üzerinden yürütülen analizler sonucu elde edilmiş ve Tablo 11'de sunulmuştur.

Tablo 11. *Matematik testindeki sorunlu maddelerin madde tepki kuramı'na göre belirlenmesine ilişkin bulgular*

	Ayırt Edicilik Bakımından	Tahmin Olasılığı Bakımından	Açıklanan Bilgi Bakımından
Sorunlu Madde Kümeleri	-	1, 2, 3, 5, 10, 11, 14, 15, 16, 19, 21, 22, 23	4, 6, 7, 9, 15, 22

Ayırt ediciliği ve taşıdığı bilgi yüksek olan bunun yanı sıra tahminle doğru yanıtlanma olasılığı çok yüksek olmayan maddeler kaliteli olarak nitelendirilebilir. Madde bilgi fonksiyonları maddelerin ayırt edicilik ve tahminle doğru yanıtlanma olasılığı gibi bilgilerinden türetilen bir fonksiyon olmasına karşın bu değerlendirmede ayrı bir kıstas olarak ele alınmış ve parametre değerlerinin bir sağlaması olarak düşünülmüştür.

MTK'ya dayalı yürütülen bu inceleme sonrasında sorunlu olduğu belirlenen 3, 4, 5, 6, 7, 10, 11, 14,16, 19, 22 ve 23. maddelerin Sato Test Kuramı ile yapılan belirlemelerle tutarlılık gösterdiği (Tablo 10) görülmüştür. Bununla birlikte 3, 4, 5, 7, 10, 11, 14, 16, 22 ve 23. maddelerin Klasik Test Kuramı ile yapılan incelemelerde de sorunlu olarak belirlendiği (Tablo 9) görülmüştür. Testteki sorunlu maddelerin her üç kurama dayalı olarak belirlenmesi ve karşılaştırılması noktasında elde edilen bulgulardan hareketle STK'nın hem KTK hem de MTK ile tutarlı sonuçlar ürettiği yorumu yapılabilir. Yapılan belirlemelerin farklı büyüklükteki örneklemler arasındaki tutarlılığı söz konusu olduğunda ise STK'nın KTK'dan daha tutarlı sonuçlar ürettiği gözlemlenmiştir.

Sonuç ve Tartışma

Araştırmanın ilk üç sorusunun STK ile KTK'nın karşılaştırılması çerçevesinde ele alındığı belirtilebilir. Bu bağlamda, incelenen testte yer alan maddelerin ayırt edicilik düzeylerinin belirlenebilmesi bakımından madde ayırıcılık (KTK) ve madde uyarı (STK) indekslerinin benzer şekilde çalıştığı sonucuna ulaşılmıştır. Ayrıca söz konusu maddelerin güçlük düzeylerinin belirlenebilmesi bakımından da madde güçlük indeksleri (KTK) ile maddeler için sınırlandırılmış gri ilişki derecelerinin (STK) benzer şekilde çalıştığı belirlenmiştir. Bireylerin test üzerinde gösterdikleri performansların belirlenmesi noktasında da toplam test puanları (KTK) ile öğrenciler için sınırlandırılmış gri ilişki dereceleri (STK) arasında tüm örneklem büyüklükleri için pozitif yönde ve yüksek düzeyde korelasyonların olduğu tespit edilmiştir. Elde edilen bulgulardan hareketle STK ile KTK'nın psikometrik özellikler bakımından birbirine yakın sonuçlar ürettiği ifade edilebilir.

Araştırmanın dördüncü, beşinci ve altıncı sorularının STK ile MTK'nın karşılaştırılması çerçevesinde ele alındığı belirtilebilir. Bu bağlamda, testte yer alan maddelerin ayırt edicilik düzeylerinin belirlenebilmesi bakımından madde uyarı indeksleri (STK) ile *a* parametreleri (MTK) arasında yüksek düzeyde bir benzerliğin bulunmadığı belirlenmiştir. Buna karşın testlerde yer alan maddelerin güçlük düzeylerinin belirlenmesi noktasında maddeler için sınırlandırılmış gri ilişki dereceleri (STK) ile *b* parametrelerinin (MTK) benzer şekilde çalıştığı sonucuna ulaşılmıştır. Bireylerin test üzerinde gösterdikleri performansların belirlenmesi söz konusu olduğunda ise öğrenciler için

sınırlandırılmış gri ilişki dereceleri (STK) ile θ değerleri (MTK) arasında pozitif yönde ve yüksek düzeyde korelasyonların bulunduğu görülmüştür. Elde edilen bulgulardan hareketle STK ile MTK'nın madde ayırt ediciliklerinin belirlenmesi haricinde benzer sonuçlar ürettiği iddia edilebilir. Bununla birlikte, STK'da doğrudan madde ayırt ediciliklerine ilişkin bilgi elde etmeye yönelik olarak kullanılabilecek bir indeksin tanımlanmadığı vurgulanabilir. Maddelerin ayırt edicilik özelliklerine işaret eden madde uyarı indeksleri tek başlarına yorumlanmamakta maddelerin doğru yanıtlanma oranlarıyla birlikte işlerlikleri yönünden sınıflandırılması amacıyla kullanılmaktadır. Araştırma bulguları, STK'nın madde uyarı indeksleriyle KTK'nın madde ayırıcılık indekslerinin benzer sonuçlar ürettiğini buna karşın MTK'nın a parametreleri ile bu benzerliğin kurulamadığını göstermektedir.

Testte yer alan sorunlu maddelerin her üç kurama dayalı olarak belirlenmesi ve karşılaştırılması sonrasında STK'ya göre sorunlu olduğu belirlenen maddelerden oluşturulan kümeler ile KTK ve MTK'ya göre sorunlu olduğu belirlenen maddelerden oluşturulan kümelerin çok sayıda ortak elemanının olduğu belirlenmiştir. Söz konusu bulgudan hareketle, sorunlu maddelerin belirlenmesi konusunda STK'ya dayalı belirlemelerin diğer kuramlara dayalı belirlemelerle paralellik gösterdiği (dış tutarlılık) sonucuna ulaşılmıştır. Ayrıca, yapılan belirlemelerin farklı büyüklükteki örneklemler arasındaki tutarlılığı söz konusu olduğunda (iç tutarlılık) STK'nın KTK'dan daha tutarlı sonuçlar ürettiği gözlemlenmiştir. Elde edilen bulgular, geçerliği ve güvenirliği yüksek ölçmelerin yapılmasına olanak tanıyacak ölçme araçlarının geliştirilmesi sürecinde STK'dan yararlanılabileceği görüşünü destekler niteliktedir.

Küçük gruplar üzerinde yürütülen ölçme ve değerlendirme uygulamalarında KTK'nın yaygın olarak kullanıldığı bilinmektedir. Bu araştırmada KTK ve STK'nın psikometrik özellikler bakımından birbirine yakın sonuçlar ürettiği ortaya konulmuş olsa da STK'yı üstün kılan, maddeler ve öğrenciler hakkında nitel dönütler elde edilebilmesi ve aynı puanı alan öğrencilerin sınıflandırmalar aracılığıyla ayırt edilebilmesi gibi özellikler dikkate alındığında STK, özellikle okullarda yürütülen ölçme ve değerlendirme uygulamaları için öğretmenlere önerilebilir. Bununla birlikte, maddelerin ayırt ediciliklerinin belirlenmesi bakımından STK ve MTK'nın farklılaştığı göz önünde bulundurulduğunda STK, özellikle MTK modellerinden 1PLM'nin tercih edileceği durumlar için de alternatif oluşturabilir.

Bu çalışmada yapılan karşılaştırmalar yalnızca iki kategorili (dichotomous) puanlanan maddelere uygun STK modeli üzerinden yürütülmüştür. Konuya ilgi duyan araştırmacılara, çok kategorili (polytomous) puanlanan maddelere uygun STK modellerinin diğer test kuramlarıyla psikometrik açıdan karşılaştırılmasını amaçlayan çalışmalar yapmaları önerilebilir.

http://kefad.ahievran.edu.tr

Ahi Evran University
Journal of Kırşehir Education Faculty

ISSN: 2147 - 1037

ENGLISH VERSION

Introduction

It can be mentioned about two major test theories, which are frequently used in order to develop tests and to analyze test results. These are named as Classical Test Theory and Item Response Theory. In the literature, it is claimed that both of these theories have superior and weak aspects when compared to each other. At this point, it is necessary to conduct research on new test theories in order to make more effective measurements in small samples and to overcome the problems which are encountered in the applications.

In the 1970s, a new technique called Student-Problem Chart Analysis was created by Japanese researcher named Takahiro Sato. It was stated that the coefficients obtained as a result of the analysis could serve as a reference for teachers to diagnose students' performances and then to guide students and improve teaching. It has been suggested that the use of this analysis is an extremely good and effective approach for formative assessment of classes with fewer students (Takeya, 1980; Tatsuoka, 1984).

Student-problem Chart Analysis was strengthened with Gray Relational Analysis in year 2010 with the suggestion of Nagai and it made usable in order to determine the ability levels of the students and their ormative assessment. Gray Relational Analysis proposes statistical solutions in decision-making situations with limited data that cannot provide the normal distribution assumption and do not contain sufficient data (Deng, 1982). This situation makes the approach, that is now known as Sato Test Theory (STT), become advantageous in the measurements made on small samples.

Item Caution Indices, Student Caution Indices, Localized Gray Relational Grade-Student and Localized Gray Relational Grade-Problem values are calculated within the scope of STT. **Item Caution Indices (ICI)** is used to assign items to four different classes. The indices takes values between 0 and 1. The fact that the ICI value of the items is higher than 0.50 provides information that there may be some problems in the related item. The ICI value is interpreted together with the correct response rate of the item.

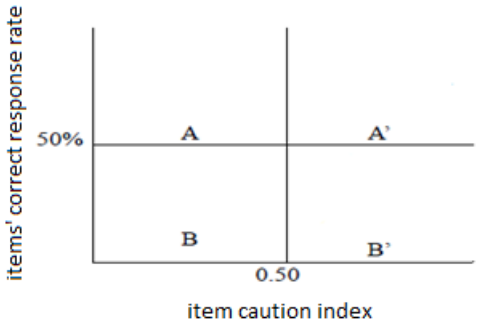


Figure 1. Classes to which items will be assigned according to ICI and ratio values.

For example, if the ICI value is less than 0.50 and the correct response rate of the item is higher than 50%, the item is assigned to Class A. The criteria for assigning items to other classes are also shown in Figure 1. The definitions that can be made regarding these classifications are given below (Sheu, Pham, Nguyen, and Nguyen, 2013).

A: Item that works well.

A': Partially problematic item. It can be reviewed.

B: It is a diffucult item.

B': A problematic item. It should be kicked out of the test or it should be rewritten.

Student Caution Indices (SCI) is used to assign students to six different classes. The indices takes a value between 0 and 1. When the value of SCI is higher than 0.50, this gives information that the relevant student may have learning deficiencies or carelessness. SCI value is interpreted together with the ratio of the number of items that the student answered correctly to the number of items in the test.

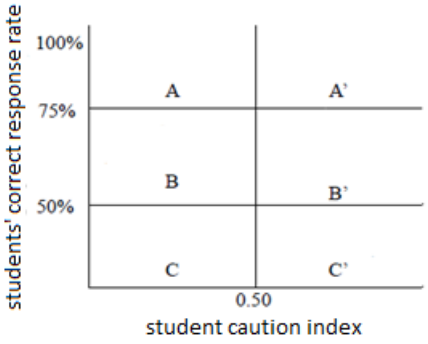


Figure 2. Classes to which students will be assigned according to SCI and ratio values.

The classifications made in line with the values in Figure 2 and the definitions that can be made about them are given below (Sheu et al., 2013).

A: Effective learning has taken place.

A': The learning has taken place, but the student is too careless.

B: The student is doing well but needs some more work.

B': The student is a little careless and needs more work.

A: The level of learning is insufficient.

C': Learning has not taken place.

Localized Gray Relational Grade-Student (LGRG-S) values provide information on the ability or success levels of students. Table 1 shows how the students' levels in the measured trait can be evaluated based on the LGRG-S values.

Table 1. *Evaluation criteria for LGRG-S values*

Student Level	LGRG-S Value*
High	LGRG-S ≥ 0,75
Upper-intermediate	0,75 > LGRG-S ≥ 0,50
Intermediate	0,50 > LGRG-S ≥ 0,25
Low	0,25 > LGRG-S

*LGRG-S value ranges from 0 to 1.

Localized Gray Relational Grade- Problem (LGRG-P) values provide information about the difficulties of the items. LGRG-P value ranges from 0 to 1, whereas high values denote easy items and low values denote difficult items. In other words, it can be said that the higher the LGRG-P value, the easier the item is.

Studies on STT have increased in recent years, especially in the Far East Geography, and the following topics have generally been emphasized in the English publications. These topics are related with the introduction of the theory (Lin and Chen, 2006; Wang, Sheu and Nagai, 2011; Wang and Chen 2013), the introduction of the software developed to apply the theory (Pham, Sheu, and Nagai, 2015; Sheu et al., 2013; Sheu et al., 2014b; Sheu et al., 2014a; Wu, 1999), the identification of students who have learned with misconceptions and/or items that cause misconceptions through STT (Sheu et al., 2013; Tsai et al., 2014) and comparison of STT and IRT (Sheu et al., 2014b; Tatsuoka, 1984) . Based on the literature review, it can be inferred that even after combining Gray Relationship Analysis with STT (2010), interest in the theory remained weak in Europe and America. It was thought that this study can contribute to the relevant literature in terms of showing the similarities and differences between different test theories and revealing the strengths and weaknesses of the proposed techniques.

Method

In this study, the item and test statistics of the mathematics subtest of the Determination of Students' Achievement Exam were obtained by using Sato Test Theory (STT), Classical Test Theory (CTT) and Item Response Theory (IRT) techniques. In this study, it was aimed to determine the relationship levels between item characteristics, to determine the problematic items that were decided

to be removed from the tests, and to compare student achievement scores which are determined by different techniques.

According to this main purpose, we sought answers to the following questions.

1. What is the correlation between Sato Test Theory's item caution indices and Classical Test Theory's item discrimination indices (biserial correlation coefficients)?

2. What is the correlation between Sato Test Theory's localized gray relational grade- problem values and Classical Test Theory's item difficulty indices?

3. What is the correlation between Sato Test Theory's localized gray relational grade-student values and Classical Test Theory's student achievement scores (total test scores)?

4. What is the correlation between Sato Test Theory's item caution indices and Item Response Theory's *a* parameters?

5. What is the correlations between Sato Test Theory's localized gray relation grade-problem values and Item Response Theory's *b* parameters?

6. What is the correlation between Sato Test Theory's localized gray relation grade- student values and Item Response Theory's ability parameters?

7. To what extent do the items that are decided to be corrected or removed from the test based on the results of the item analyzes with Classical Test Theory, Item Response Theory and Sato Test Theory agree?

Dataset

The research dataset consists of the responses of 5461 8th grade students who took the Determination of Students' Achievement Exam (DSAE) applied in year 2005. The dataset was created with binary scored (1-0) data.

There were 25 items with four options in the mathematics subtest of DSAE, which was evaluated within the scope of this research. The histogram graph of the scores of the students who took the exam in the math test is given in Figure 3.

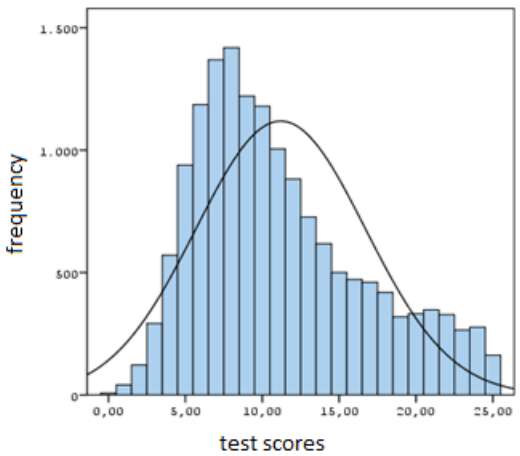


Figure 3. Histogram plot of the distribution of math test scores.

The mean of the distribution of test scores in Figure 3 was 11,199, the median was 10,000, and the standard deviation was 5.513. The skewness coefficient of the distribution, which appeared to right-skewed, was 0.697. The kurtosis coefficient of the distribution was -0.320.

Testing Assumptions and Model Selection

At this stage, firstly, the unidimensionality assumption required for both CTT and IRT was examined with the Parallel Analysis method. The mentioned method is described by many researchers as the best method for determining the number of factors (Fabrigar, Wegener, MacCallum, and Strahan, 1999; Field, 2009; Hayton, Allen, and Scarpello, 2004; Henson and Roberts, 2006; Horn, 1965; Thompson, 2004; Zwick and Velicer, 1986). For this purpose, factor analysis based on tetrachoric correlation matrix was carried out and the Unweighted Least Squares method proposed by Coughlin (2013) was chosen as the factorization method. Evidence for the factorability of the correlation matrix was obtained with the Bartlett's Test of Sphericity and the Kaiser-Meyer-Olkin (KMO) Test of Sampling Adequacy. The results of the Bartlett Test of Sphericity indicate that there was a statistically significant difference (p= 0.000) between the correlation matrix and the unit matrix. In addition, considering the value (0.901) obtained from the KMO test, it could be concluded that the factorability levels of the correlation matrices were "excellent" (Beavers et al., 2013). The results of the Parallel Analysis performed after providing the assumptions regarding the correlation matrix suggested "one dimension" for the test. Thus, we decided that the research data met the unidimensionality assumption. Unidimensionality can also be shown as evidence for the existence of local independence, which is another IRT assumption. In addition to this, the standardized chi-square values between the items were examined and it was concluded that the inter-item relations did not violate local independence.

After checking the assumptions, the IRT model that best fitted the data was selected so that IRT parameters could be estimated more accurately. For this purpose, the test data were analyzed according to 1P, 2P and 3P Logistic Models and the differences between the -2log Likelihood values were examined. These values were calculated as 471051,75 for 1PLM, 468898,42 for 2PLM, and 463975,82 for 3PLM.

The χ^2 difference test was used to compare the fit levels of the models. It was concluded that the most fit model for the test was the 3PL Model. The calculations made are given below.

$\chi^2 = -2\log Likelihood_{1plm} - (-2\log Likelihood_{2plm}) = 471051{,}75 - 468898{,}42 = 2153{,}33 > 37{,}65\ (\chi^2_{(df=25\ p=;0{,}05)})$

$\chi^2 = -2\log Likelihood_{2plm} - (-2\log Likelihood_{3plm}) = 468898{,}42 - 463975{,}82 = 4922{,}60 > 37{,}65\ (\chi^2_{(df=25\ p=;0{,}05)})$

We examined the invariance of the item and ability parameters produced with 3PLM in order to gain further evidence of the fit of the selected model to the data. In order to conduct this, we randomly divided the response patterns of individuals who took the math test into two groups and examined the relationships between the item parameters estimated from both groups with Spearman-Brown Rank Differences Correlation Coefficients. The correlation coefficients between the item parameters from the two groups were 0.926 for the *a* parameters; 0.993 for *b* parameters; 0.940 for the *c* parameters. Based on this finding, it could be interpreted that item parameters show invariance in 3PLM.

In order to examine the invariance of ability parameters, we created subtests by randomly dividing the items in the math test into two groups, and we estimated the ability parameters from these subtests with 3PLM. We analyzed the correlation level between the ability parameters estimated over two separate subtests with the Pearson Correlation Coefficient technique and found that this value was 0.887. Based on the aforementioned finding, it could be claimed that the ability levels of individuals who took the test were independent of the item sample in the test. After this process, we decided to use the Three-Parameter Logistic Model that best fits the research data in the analysis and to limit comparisons with other test theories to the parameters of this model.

Data Analysis Process

First of all, the research dataset was prepared to analyze the research questions by randomly sampling the math test response forms of 15461 8th grade students who participated in DSAE (2005), by creating samples of 50, 100, 200 and 600, 10 of each.

In order to find an answer to the first question of the research, r_{jx} and ICI values of the test items were obtained from all the samples created. In order to answer to the second question of the study, p and LGRG-P values were obtained. The correlation coefficients between the mentioned values were calculated with Spearman's Rank Difference Correlation Coefficient Technique over 10 different samples of each sample size (50, 100, 200 and 600) and their medians were taken.

In order to seek an answer to the third question of the study, the test total scores and LGRG-S values of the students in the samples were obtained. The correlation coefficients between the mentioned values were calculated with the Pearson Correlation Coefficient Technique on 10 different samples of each sample size and their medians were taken.

In order to seek answers to the fourth and fifth questions of the study, ICI and LGRG-P values were obtained from 10 different samples of each sample sizes for the items in the test, and a and b parameters were obtained with 3PLM over 15461 students (whole dataset). The correlation coefficients between these variables were calculated using Spearman's Rank Differences Correlation Coefficient Technique and their medians were taken.

In order to find an answer to the sixth question of the study, the correlations between the LGRG-S values of the students who took the test and their ability (θ) parameters estimated by the 3PL IRT model were examined using the Pearson Correlation Coefficient Technique.

In order to compare the problematic items, the CTT discrimination indices were calculated using the Biserial Correlation Coefficient technique. Then, the same data were analyzed based on Sato Test Theory and the problematic items were determined through the classifications made with ICI's. Finally, the problematic items with the a, c parameters and item information functions of the Item Response Theory were determined.

The MATLAB Package developed by Sheu, Pham, Nguyen, and Nguyen (2013) was used for the Sato Test Theory estimates in the analysis of the data. IRTPRO (V, 4.2) software was used for Item Response Theory estimates and TAP (V, 14.7.4) software was used for Classical Test Theory statistics. SPSS (V, 22) software was used for Spearman's Rank Differences Correlation Coefficient Technique, Pearson's Correlation Coefficient Technique and Kruskal-Wallis-H Test.

Findings

The correlation coefficients we calculated to answer to the question of "What is the correlation between Sato Test Theory's item caution indices and Classical Test Theory's item discrimination indices?" are given in Table 2.

Table 2. *Findings on the correlations between item discrimination indices and item caution indices*

Sample Size	50	100	200	600
$\mathbf{Corr}_{(ICI,r_{jx})}$	-0,968*	-0,934*	-0,944*	-0,934*

*The values given were obtained by finding the medians of the correlation coefficients obtained from 10 different samples of each sample size.

When Table 3 was examined, it was seen that there were negative and high level correlations between the item discrimination indices of the CTT and the item caution indices of the STT for all sample sizes. Based on the high correlation coefficients obtained, it can be argued that item discrimination and item caution indices work similarly in terms of determining the discrimination levels of the items in the test. Additionally, there was no finding that correlations were affected by sample size.

The correlation coefficients we calculated to answer to the question of "What is the correlation between Sato Test Theory's localized gray relational grade- problem values and Classical Test Theory's item difficulty indices?" are given in Table 3.

Table 3. *Findings on correlations between item difficulty indices and localized gray relation grades-problems*

Sample Size	50	100	200	600
$\mathbf{Corr}_{(\mathbf{LGRG-P},p)}$	1,000*	0,998*	0,998*	0,999*

*The values given were obtained by finding the medians of the correlation coefficients obtained from 10 different samples of each sample size.

When Table 3 was examined, it was seen that there were positive and high correlations between the item difficulty indices of the CTT and the localized gray relation grades-problem of the STT in all sample sizes. There was no finding that the correlations were affected by the sample size. Based on the high correlation coefficients obtained, it can be argued that item difficulty indices and LGRG-P work similarly in terms of determining the difficulty levels of the items in the test.

The correlation coefficients we calculated to answer to the question of "What is the correlation between Sato Test Theory's localized gray relational grade-student values and Classical Test Theory's student achievement scores (test total scores)?" is given in Table 4.

Table 4. *Findings on the correlations between test total scores and localized gray relation grades- students*

Sample Size	50	100	200	600
$\mathbf{Corr}_{(\mathbf{LGRG-S,TTS})}$	0,979*	0,977*	0,978*	0,976*

*The values given were obtained by finding the medians of the correlation coefficients obtained from 10 different samples of each sample size.

When Table 4 was examined, it was seen that there were positive and high-level correlations between the test total scores and the localized gray relation grades- students in all sample size. In addition to this, there was no finding that correlations were affected by sample size. It can be stated that LGRG-S and test total scores are similar in determining the performance of individuals on the test. When the findings related to the first three questions of the study were evaluated, it was seen that the STT and CTT revealed similar results.

The correlation coefficients we calculated to answer to the question of "What is the correlation between Sato Test Theory's item caution indices and Item Response Theory's a parameters?" are given in Table 5.

Table 5. *Findings on the correlations between item caution indices and a parameters*

Sample Size	50	100	200	600
$\mathbf{Corr}_{(\mathbf{ICI,a})}$	-0,082*	-0,154*	-0,138*	-0,269*

*The values given were obtained by finding the medians of the correlation coefficients obtained from 10 different samples of each sample size.

When Table 5 was examined, it was seen that there was a negative and low-level correlation between the item caution indices of the STT and the a parameters of the IRT in all sample sizes. In this respect, it can be stated that ICI and a parameters do not work similarly in determining the discrimination of test items. Note that *a* parameters are estimated with the 3PL model using the total dataset.

The correlation coefficients we calculated to the question of "What is the correlations between Sato Test Theory's localized gray relation grade-problem values and Item Response Theory's *b* parameters?" are given in Table 6.

Table 6. *Findings on the correlations between localized gray relation grades- problems and b parameters*

Sample Size	50	100	200	600
$\mathbf{Corr}_{(\mathbf{LGRG-P},b)}$	-0,805*	-0,880*	-0,891*	-0,928*

*The values given were obtained by finding the medians of the correlation coefficients obtained from 10 different samples of each sample size.

When Table 6 was examined, it was seen that there were negative and high correlations for all sample sizes between the localized gray relation grade-problems of the STT and the *b* parameters of the IRT. In addition, it was determined that the correlation coefficients between LGRG-P values estimated from larger samples and *b* parameters were greater. It should be noted that the *b* parameters were estimated from the whole dataset at all time. Based on this findings, it can be stated that LGRG-P and *b* parameters are similar in determining the difficulty levels of test items.

The correlation coefficients we calculated to answer to the question of "What is the correlation between Sato Test Theory's localized gray relation grade- student values and Item Response Theory's ability parameters?" are given in Table 7.

Table 7. *Findings on the correlations between students' ability parameters and localized gray relation grades-students*

Sample Size	50	100	200	600
$\mathbf{Corr}_{(\mathbf{LGRG-S},\boldsymbol{\theta})}$	0,936	0,915	0,941	0,949

In Table 7, it was seen that there were positive and high correlations for all sample sizes between the localized gray relation grades-students of STT and the ability parameters of IRT. There was no clear finding that correlations were affected by the size of the samples from which LGRG-S values were obtained. Based on the findings, it can be stated that LGRG-S and θ values are similar in determining the performance of individuals on the test.

To what extent do the items that are decided to be corrected or removed from the test based on the results of the item analyzes with Classical Test Theory, Item Response Theory and Sato Test Theory agree? In order to seek an answer to this question, in the first stage, item discrimination indices (r_{jx}) were calculated using the biserial correlation coefficient technique of the items in the math test. Based on these values, the problematic items (r_{jx}<0.30) that were decided to be corrected or removed from the test were detected and given in Table 8 for each sample size.

Table 8. *Determination of problematic items in mathematics test according to classical test theory*

Sample Size	50	100	200	600
Problematic Item Sets	3, 5, 6, 11, 13, 14, 16, 22, 23	4, 6, 7, 16	10, 22, 23	4, 6, 7, 16, 23

As Table 8 shows, items 6, 16 and 23 were the most repeated items found to be problematic in different sample sizes according to the CTT.

For the examinations based on the Sato Test Theory, the items were classified via ICI values and correct response rates. As a result of the examinations, the problematic items that were decided to be corrected or removed from the test were detected and given in Table 9 for each sample size.

Table 9. *Determination of problematic items in mathematics test according to sato test theory*

Sample Size	50	100	200	600
Problematic Item Sets	3, 5, 6, 11, 13, 14, 16, 19, 22, 23	4, 6, 7, 10, 13, 16, 22, 23, 25	7, 10, 14, 22, 23	4, 7, 10,16, 22, 23

As seen in Table 9, items 22 and 23 were classified as problematic in all sample sizes, and items 7 and 10 were classified as problematic in samples of 100, 200 and 600. When the item sets found to be problematic after the analyzes based on STT and CTT were compared, while the sample size was 50, items 3, 5, 6, 11, 13, 14, 16, 22 and 23, while it was 100, items 4, 6, 7 and 16, while it was 200, items 10, 22 and 23, while it was 600, items 4, 7, 16 and 23 were common. Based on the comparisons made, it can be claimed that the analyzes carried out for the detection of problematic items based on both theories reveal similar results.

Problematic items that need to be corrected or removed from the test with Item Response Theory were determined through a, c parameters and item information functions. The whole dataset that was used for the analyzes and the findings are given in Table 11.

Table 10. *Determination of problematic items in mathematics test according to item response theory*

	In terms of discrimination	In terms of guessing	In terms of information
Problematic Item Sets	-	1, 2, 3, 5, 10, 11, 14, 15, 16, 19, 21, 22, 23	4, 6, 7, 9, 15, 22

Items with high discriminative power, high level of information and not very high probability of being answered correctly by guessing can be qualified as high quality. After this examination based on IRT (Table 10), items 3, 4, 5, 6, 7, 10, 11, 14, 16, 19, 22, and 23, which were found to be problematic, were found to be consistent with the detection made with the STT (Table 10). Additionally, items 3, 4, 5, 7, 10, 11, 14, 16, 22, and 23 were also found to be problematic in the detection made with CTT (Table 9). According to the findings obtained from the detection and comparison of the problematic items based on all three theories, it can be interpreted that STT produces results consistent with both CTT and IRT. When it comes to the consistency of the detections between different sized samples, it was observed that STT produces more consistent results than CTT.

Conclusion and Discussion

It can be stated that the first three questions of the research were handled within the framework of the comparison of STT and CTT. It was concluded that item discrimination (CTT) and item caution (STT) indices work similarly in terms of determining the discrimination levels of the items in the examined test. In addition, it was determined that item difficulty indices (CTT) and localized gray relation grades-problems (STT) work similarly in terms of determining the difficulty levels of the items. At the point of determining the level of the students' performances on the test, it was seen that there were positive and high level correlations for all sample sizes between the total test scores and the localized gray relation grades- students. Based on the findings, it can be stated that STT and CTT produce similar results in terms of psychometric examinations.

It can be stated that the fourth, fifth and sixth questions of the research were evaluated within the framework of the comparison of STT and IRT. It was determined that there was not a high level of similarity between the item caution indices (STT) and a parameters (IRT) in terms of determining the discrimination levels of the items in the test. On the other hand, it was concluded that the localized gray relation grades-problems (STT) and b parameters (IRT) worked similarly at the point of determining the difficulty levels of the items in the tests. When it came to determining the performance level of students, it was seen that there were positive and high level correlations between localized gray relation grades-students (STT) and θ values (IRT). Based on the findings, it can be claimed that STT and IRT produce similar results, except for the determination of item discrimination. However, it can be emphasized that there is no indices that gives information about item discrimination in STT. Item caution indices, which are assumed to indicate the discrimination of the items, are not interpreted alone, but are used to classify the items in terms of their functionality with items' correct response rates. The findings show that the item caution indices of the STT and the item discrimination indices of the CTT produce similar results, but this similarity cannot be established with the a parameters of the IRT.

After the comparison of the problematic items in the test based on all three theories, it was determined that the sets formed from the problematic items according to the STT and the sets formed from the problematic items according to the CTT and the IRT had many common elements (items). Finally, it was determined that the internal consistency of the STT estimates was high in terms of invariance in samples of different sizes. The findings of this research support the view that the STT can be used in the process of developing measurement tools (tests) that allow measurements with high validity and reliability.

It is known that CTT is widely used in assessments carried out on small groups. Although it was revealed in this study that CTT and STT produced similar results in terms of psychometric properties, there are some features that make STT become advantageous. Considering the situations

such as obtaining qualitative feedback about items and students, and distinguishing students with the same score through classifications made with STT, it can be recommended to teachers especially for assessment practices carried out in schools. STT is also more useful for online exams applied with the aim of giving quick and functional feedback to individuals/students. Additionally, considering that STT and IRT differ in terms of determining the discrimination levels of items, STT can also create an alternative especially for cases where 1PL IRT model would be preferred. Comparisons made in this study were limited to the STT model only for items scored in two categories (dichotomous). We can suggest other researchers interested in the subject to conduct studies aiming at psychometric comparison of STT models used in polytomous scored items with other test theories.

Kaynakça

Beavers, A.S., Lounsbury, J.W., Richards, J.K., Huck, S.W., Skolits, G.J. & Esquivel, .L. (2013). Practical considerations for using exploratory factor analysis in educational research. *Practical Assessment, Research and Evaluation, 18*(6), 1-13.

Coughlin, K.B. (2013). *An analysis of factor extraction strategies: A study of the relative strenghts of principal axis, ordinary least squares and maximum likelihood factor extraction methods in research contexts*. Unpublished Doctoral Dissertation. University of South Florida, Tampa, FL, United States of America.

Çüm, S., Gelbal, S. & Tsai, C.P. (2016). Sato test kuramı yöntemleriyle farklı örneklemlerden elde edilen madde parametrelerinin tutarlılığının incelenmesi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 7*(1),170-181.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: The Guilford Press.

Deng, J.L. (1982), Control problems of grey systems, *Systems & Control Letters, 1*(5), 288-94.

Erkuş, A. (2010). Psikometrik terimlerin Türkçe karşılıklarının anlamları ile yapılan işlemlerin uyuşmazlığı. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi,1*(2), 72-77.

Fabrigar, L. R., Wegener,D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*(3), 272-299.

Field, A. (2009). *Discovering statistics using SPSS.* London: Sage Publications Ltd.

Hayton, J.C., Allen, D.G. & Scarpello, V. (2004). Factor retention decision in exploratoy factor analysis: A tutorial on parallel analysis. *Organizational Research Methods, 7*(2), 191-205.

Henson, R.K. & Roberts, J.K. (2006). Use of exploratory factor analysis in published research: common errors and some comment on improved practice. *Educational and Psychological Measurement, 66*, 393-416.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*(2),179-85.

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

Lin, Y.H., & Chen, S.M. (2006). *The investigation of S-P chart analysis on the test evaluations of equality axiom concepts for sixth graders*. Proceedings of the 2nd International Conference on Educational Technologies, Romania, Bucharest.

Lin, Y.H., & Yih, J.M. (2015). *Application of IIRS in mathematics instruction to promote pupils decimal concept*. The International Conference on Language, Education and Psychology, Taiwan.

Pham, D.H., Sheu, T.W., & Nagai, M. (2015). PCSP 1.0 software for partial credit S-P chart analysis. *International Journal of Hybrid Information Technology*,*8*(6), 309-322.

Ree, M. J., & Jensen, H. E. (1983). *Effects of sample size on linear equating of item characteristic curve parameters: Latent trait test theory and computerized adaptive testing.* New York: Academic Press.

Sheu, T. W., Nguyen, P. T., Nguyen, P. H., Pham, D. H., Tsai, P. C., & Nagai, M. (2014b). A MATLAB toolbox for misconceptions analysis based on S-P chart, grey relational analysis and ROC. *Transactions on Machine Learning and Artificial Intelligence*, *2*, 72-85.

Sheu, T.W., Nguyen, P.T., Tsai, C.P., Pham, D.H., Nguyen, P.H. & Nagai, M. (2014c). Using grey student-problem chart in the evaluation of tests with large data sets. *Education Practise and Innovation*, 1(2), 2372-3106.

Sheu, T.W., Pham, D.H., Nguyen, P.T., & Nguyen, P.H. (2013). Amatlab toolbox for student-problem chart and grey student-problem chart and its application. *International Journal of Kansei*, *4*(2), 75-86.

Sheu, T.W., Pham, D.H., Tsai, C.P., Nguyen, P.T., Nguyen, P. H. & Nagai, M. (2014a). Rasch GSP toolbox for assessing academic achievement. *Journal of Software*, *9*(7), *1903-1913.*

Sheu, T.W., Tsai, C.P., Tzeng, J.W., Pham, D.H., Chiang, H.J., Chang, C.L. & Nagai, M. (2013). An improved teaching strategies proposal based on student's learning misconceptions. *International Journal of Kansei Information, 4*(1), *1-12.*

Takeya, M. (1980). Construction and utilization of item relational structure graphs for use in test analysis. *Japan Journal of Educational Technology*, *5*, 93-103.

Tatsuoka, K. K. (1984). Caution indices based on item response theory. *Psychometrika, 49*(1), 95-110.

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: understanding concepts and applications*. Washington, DC: American Psychological Association.

Tsai, C.P., Sheu, T.W., Tzeng, J.W.,Chen, H.J., Chiang, H.J. & Nagai, M. (2014). Diagnose learning misconceptions based on rough sets. *International Journal of Applied Mathematics and Statistics, 52(2*), 63-75.

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

Wang, B.T., Sheu, T.W., & Nagai, M. (2011). Evaluating the english-learning of engineering students using the grey S-P chart: a facebook case study in Taiwan. *Global Journal of Engineering Education*, *13*(2), 51-56.

Wang, C.H. & Chen, C.P. (2013). Employing online S-P diagnostic table for qualitative comments on test results. *The Electronic Journal of e-Learning*, *11*(3), 263-271.

Wu, H. (1999). Software Based on S-P Chart Analysis and Its Applications. *Proceedings of the National Science Council*, *8,* 102-107.

Zwick,W. R. & Velicer,W. F. (1986). Factors influencing five rules for determining the number of components to retain. *Psychological Bulletin, 99,* 432-44.