ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2018; 9(2);178-201

Toplam Test ve Alt Test Puanlarının Kestiriminin Hiyerarşik Madde Tepki Kuramı Modelleri ile Karşılaştırılması*

Comparison of Estimation of Total Score and Subscores with Hierarchical Item Response Theory Models

Sümeyra SOYSAL ** Hülya KELECİOĞLU ***

Öz

Bu araştırmada güvenilir alt test ve toplam test puanı kestirimleri konusuna katkı sağlamak amacıyla alt test ve toplam test arasındaki ilişki hiyerarşik madde tepki kuramı modelleri ile araştırılmak istenmiştir. Çalışmada Üst Düzey Sıralı (Higher Order), İki Faktör (Bi-factor) ve hiyerarşik çok boyutlu madde tepki kuramı (ÇBMTK) modelleri ile kestirilen toplam test puanının ve alt test puanlarının RMSE ve güvenirlik değerleri alt test sayısı, alt test uzunluğu ve alt testler arasındaki korelasyonların büyüklüğü koşulları altında karşılaştırılmıştır. Ayrıca TEOG 2015 verileri üzerinde araştırmada kullanılan üç kestirim modelinin performansı incelenmiştir. Araştırmanın sonucunda iki ve üç boyutlu verilerde hemen hemen tüm koşullarda alt test uzunluğu ve alt testler arasındaki korelasyonun arttıkça üç kestirim modelinden elde edilen toplam test puanı için yetenek parametreleri kestirim hatasının azaldığı, kestirim güvenirliğinin ise arttığı bulunmuştur. Toplam test puanları için Hiyerarşik ÇBMTK model ile tüm koşullarda en düşük RMSE değeri ve en yüksek güvenirlik değeri elde edilmiştir. Ayrıca korelasyonun 0.8 düzeyinde toplam test puanı için tüm modeller birbirine yakın RMSE ve güvenirlik değerleri ile kestirim yapmıştır. İki ve üç boyutlu verilerde alt test puanı için kestirilen yetenek parametrelerinin RMSE değerleri, Hiyerarşik ÇBMTK modelde alt test uzunluğu arttıkça azalırken alt testler arasındaki korelasyon düzeyinden etkilenmediği; Üst Düzey Sıralı modelde alt test uzunluğu ve alt testler arasındaki korelasyon arttıkça azaldığı; İki Faktör modelde ise alt test uzunluğu arttıkça azalırken alt testler arasındaki korelasyon arttıkça önemli düzeyde arttığı bulunmuştur.

Anahtar Kelimeler: Alt test puan kestirimi, toplam test puan kestirimi, hiyerarşik madde tepki kuramı modelleri, üst düzey sıralı model, iki faktör model

Abstract

In this study, the relationship between subtest and total test was investigated by using hierarchical item response theory models in order to contribute to reliable subtest and total test score estimates. The RMSE and reliability of the total test score and subtest scores estimated by the Higher Order, Bi-factor and hierarchical MIRT models in the study were compared under the conditions of the size of the correlations between the subtests, subtest length and number of subtests. In addition, the performance of three models used in the research was examined on TEOG 2015 data. As a result of the study, in almost all conditions, when the correlation between the subtest and the subtest length increased, the RMSE of the ability parameters decreased and the reliability increased for the total test score obtained from the three estimation models. Under all conditions, the lowest RMSE values and the highest reliability values were yielded from Hierarchical MIRT model for subtest score recovery and from Hierarchical MIRT model for total test score recovery. In addition, all models estimated RMSE and reliability values close to each other at 0.8 level of correlation for total test score recovery. The RMSE values of the ability parameters for the subtest scores in two and three dimensional data were found to be not affected by the correlation level between the subtests while the subtest length decreased in the Hierarchical MIRT model; were found to decrease as the correlation between subtest and subtest length in the Higher Order model and were found to decrease as the subtest length increased, but significantly increased as the correlation between the subtests increased in the Bi-factor model.

* Bu çalışma, ilk yazarın, ikinci yazar danışmanlığında tamamladığı "Toplam Test Puanı ve Alt Test Puanlarının Kestiriminin Hiyerarşik Madde Tepki Kuramı Modelleri ile Karşılaştırılması" isimli doktora tezinden üretilmiştir.

**Dr., Hacettepe Üniversitesi, Eğitim Fakültesi, Ankara-Türkiye, sumeyrasoysal@hotmail.com, ORCID ID: orcid.org/0000-0002-7304-1722

*** Prof. Dr., Hacettepe Üniversitesi, Eğitim Fakültesi, Ankara-Türkiye, hulyakelecioğlu@gmail.com, ORCID ID: orcid.org/0000-0002-0741-9934

Geliş Tarihi: 11.03.2018
DOI: 10.21031/epod.404089 *Kabul Tarihi: 13.06.2018*

Keywords: subtest scoring, overall test scoring, hierarchical item response theory models, higher order model, bi-factor model

GİRİŞ

Pek çok gelişmiş ülkede, geniş ölçekli standart testler, eğitimde ve psikolojide kullanılan en yaygın ölçme araçlarıdır. Ülkemizde bu araçlar, bir eğitim programına giriş, sertifika alımı ya da personel seçimi gibi önemli kararların verildiği durumlarda sıklıkla kullanılmaktadır. Birçok ülke geniş ölçekli testleri kendi eğitim sistemlerine yönelik bilgiler toplama, eğitime yönelik karar verme ve planlama aşamalarında sıkça kullanmaktadır. SAT ve ACT gibi Amerika'da yapılan geniş ölçekli sınavların burs verme ve eyaletlerin eğitim politikalarının değerlendirilmesi gibi ikincil amaçları vardır. Öğrenci gelişimlerinin izlenmesinde, okulların başarı durumlarının/performanslarının yıllara göre incelenmesinde, öğretim programlarının araştırılması, değerlendirilmesi ve geliştirilmesinde geniş ölçekli testlerden yararlanılmaktadır. Bu testlerin sonuçlarından öğrenciler, öğretmenler, veliler, yöneticiler ve diğer paydaşlar farklı şekillerde faydalanmaktadır.

Geniş ölçekli sınavlar, genellikle hem farklı yapıları hem de bir yapının farklı alt alanlarını ölçen alt bölümlerden oluşur. Bu alt bölümlere genellikle alt test, alt bölümlerden elde edilen puanlara da alt test puanı denir. Örneğin, KPSS sınavındaki eğitim bilimleri alt testi, kendi içinde sekiz konu alanına ayrılır (öğrenme psikolojisi, gelişim psikoloji, ölçme ve değerlendirme, rehberlik ve özel eğitim, öğretim ilke ve yöntemleri, program geliştirme, sınıf yönetimi, öğretim teknolojileri ve materyal tasarımı). Bir alt testteki maddeler bir yeteneği, bir konu alanını ya da bir örtük yapıyı ölçmek için düzenlenirler. Ülkemizde bu testlerin sonuçları alt testlerin ağırlıklı ortalamalarından elde edilen toplam puanlar ile ifade edilir. Böyle birleşik puanlar genel olarak birey başarısını değerlendirmek için yeterli bilgiyi sağlayabilirler. Çünkü geniş ölçekli testlerin en başta oluşturulma amacı bireyleri sıralamaktır.

ABD'de 2001 yılında kabul edilen Hiçbir Çocuk Geride Kalmasın (No Child Left Behind) Yasası gereğince her eyalet okul ilerleme durumunu ölçmek amacıyla toplam test puanı yanında öğrencilerin matematik, okuma, yazma, fen bilimleri gibi temel konu alanlarındaki puanlarını da raporlaması gerekmektedir. Brennan (2012) test puanlarını kullananların çoğunlukla toplam test puanıyla birlikte alt testlerin tanısal amaçlar için raporlanmasını talep ettiğini belirtmiştir. Ayrıca Haladyna ve Kramer (2005), testin başlangıçtaki temel amacı ne olursa olsun, öğrenciler, öğretmenler, veliler, yöneticiler ve diğer paydaşlar tarafından farklı alt alanlar ya da alt bölümlere ait puanların büyük talep gördüğünü raporlamışlardır (akt. Ling, 2012).

Alt testler, formatif (biçimlendirici) ve summatif (özetleyici) değerlendirmelere, eğitim programlarının değerlendirilmesine ve öğretmen değerlendirmelerine bilgi sağlayabilecek bir potansiyele sahiptir. Benzer şekilde, alt testler, toplam puanla karşılaştırıldığında, bireylerin yeteneklerinin farklı alanlarda nasıl değiştiğini/ çeşitlendiğini belirlemek için daha bilgilendirici olabilmektedir. Bilişsel bir yapıyı, bir yeteneği ya da psikolojik bir yapıyı temsil eden ve bunlara yönelik tanılayıcı bilgiler sağlayan alt testler, sınıf içi ve dışı etkinliklerin düzenlenmesinde de yararlı olabilmektedir. Testlerde başarısız olan bireyler, testin kapsamında yer alan konu alanları, yeterlik alanları ya da bilişsel yapılar içinde başarılı ve başarısız oldukları noktaları bilmek istemektedir. Böylece bireyler, çalışma planlarını eksik ya da zayıf oldukları konuları tamamlayabilmek için daha etkili şekilde düzenleyebilme imkanı bulmaktadır (Haladyna ve Kramer, 2004). Ayrıca alt testler öğrencilerin güçlü ve zayıf oldukları noktalar hakkında bilgi sağlayarak öğretmenlerin ders programlarını düzenlemesine katkı sağlayabilmektedir. Yine alt testlerin sağladığı bilgilerle veliler çocuklarının durumları ile ilgili bilgilenirken, onların eksik veya başarısız oldukları konular için destekleyici tedbirler alma ya da onların potansiyellerine göre yönlendirici imkanlar sağlama konusunda daha etkili çözüm üretebilmektedir.

Alt testlerden elde edilebilecek bilgilerden yararlanabilmek için öncelikle alt test puanlarının test geliştirme süreçleri açısından bazı önemli özellikleri sağlaması gerekmektedir. (ETS, 2014; Ferrara ve DeMauro, 2007). İlk olarak, alt test puanları güvenirlik, geçerlik, ayırt edicilik açısından yeterli

psikometrik niteliklere sahip olmalıdır. Psikolojide ve Eğitimde Test Geliştirme Standartları 5.12'ye (AERA, APA, NCME,1999) göre test puanlarının geçerliği, güvenirliği ve karşılaştırılabilirliği sağlanmadıkça raporlanmaması gerekir. Yine aynı standartların 1.12'ye göre, bir test birden fazla puan sağlıyorsa farklı puanların ayırıcılıklarının gösterilmesi gerekir. Benzer şekilde, Ferrara ve DeMauro (2007) orta düzeyde ilişkili ve yüksek güvenirliğe sahip alt testlerin raporlanmasını, düşük güvenirlikli alt testlerin ise raporlanmaması gerektiğini belirtmişlerdir. Alt testlerin güvenirliğine ek olarak testin yapısının da incelenmesi gerekir. Messick'e (1989, s.43) göre maddeler arası ilişkiler yapının alt testlerini ya da alt alanlarını yansıtmalı ve bu da test puanları ve onların yorumlanması düzeyinde ele alınmalıdır. İkinci olarak, Haberman (2008) ve Haberman, Sinharay ve Puhan (2009) alt test puanlarının toplam puan üzerinde bir değeri olup olmadığının belirlenmesi gerektiğini belirtmişlerdir.

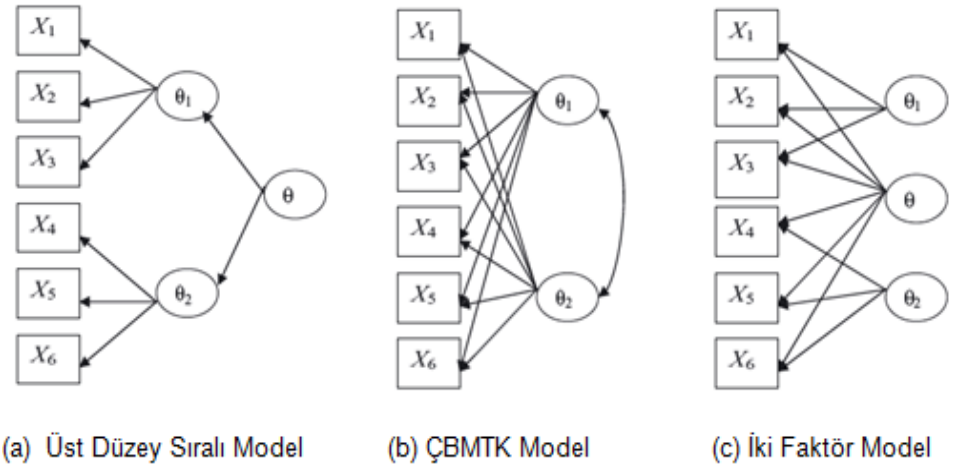
Bir testin alt birimlerinden (alt yeterlik alanları, alt testler, alt ölçekler vb) elde edilen puanların genel olarak zayıf psikometrik özelliklere sahip olduğu belirtilmiştir (Monaghan, 2006; Skorupski ve Carvajal, 2010). Alt birimler testin toplamına göre daha az sayıda madde içermesi nedeniyle daha düşük güvenirliğe sahip olabilmektedir. Sinharay (2010) yetersiz test uzunluğuna sahip olması nedeniyle alt testlerden alınan puanların güvenilir olmasa dahi potansiyel tanı değerleri nedeniyle raporlanmasının yararlı olabileceğini belirtmesine rağmen güvenilir alt test puanlarının raporlanması gerekli ve önemlidir (Haberman, 2008). Bu durumda, "Az sayıda madde içermesi nedeniyle bir testin daha küçük alt birimlerinden güvenilir puanlar elde edilebilir mi veya alt testler arası korelasyon, alt test uzunluğu gibi test özelliklerinin alt test puanı kestirimleri üzerindeki etkisi nedir?" gibi sorular ortaya çıkmaktadır.

Alt test puan ya da birleşik toplam puan kestirimlerinde geleneksel Klasik Test Kuramı'na (KTK) dayalı toplam puanı ya da düzeltilmiş toplam puanı kullanan yöntemler bulunmaktadır (Kelley, 1927, 1947; Wainer ve diğerleri, 2001). KTK'nın madde ve birey örneklemine bağlı olması nedeniyle daha güvenilir alt test puan kestirimlerinde tek boyutlu Madde Tepki Kuramı'na (MTK) dayalı yöntemler geliştirilmiştir (Wainer ve diğerleri 2001;Yen, 1987). Her ne kadar alt test puanı kestirimlerinde tek boyutlu MTK'ya dayalı yöntemlerin KTK'ya dayalı yöntemlere göre daha güvenilir sonuçlar verdiği gösterilse de tek boyutlu yaklaşımlar alt testler arası ilişkileri göz ardı etmektedir. Geniş ölçekli testlerde toplam puanların alt testlerin tek boyutlu olduğu varsayımı altında ve birbirleri arasındaki ilişkilerden bağımsız olarak analiz edilmesi durumunda eğer test maddeleri yerel bağımlı ise güvenirliğin ve test parametrelerinin yanlı kestirilmesi beklenen bir sonuçtur (Brandt ve Duckor, 2013; Wang ve Wilson, 2005;Yen, 1980). Bu sorunun çözümünde alt testlerden oluşan testler için toplam test yetenek kestirimlerinde MTK 'ya dayalı çeşitli modellerin kullanıldığı görülmektedir. Alt testlerden kaynaklanan yerel bağımlılığın, test yapısına (madde demetleri kullanımı gibi) veya ölçülen psikolojik yapıya (alt yeterlik alanları gibi) bağlı olup olmadığına göre bu modeller sırasıyla testlet modeller (Bradlow, Wainer ve Wang, 1999; Wang ve Wilson, 2005) ya da hiyerarşik modeller (de la Torre ve Song, 2009; Gibbons ve Hedeker, 1992; Sheng ve Wikle, 2008) olarak gösterilmektedir.

Alt testleri çok boyutlu MTK çerçevesinde analiz eden birkaç çalışma bulunmaktadır (de la Torre, Song ve Hong, 2011; Sheng ve Wikle, 2007; Wang, Chen ve Cheng, 2004; Yao, 2010; Yao ve Boughton, 2007). Ayrıca, ülkemizde yapılan bilimsel araştırmalar incelendiğinde çok boyutlu yetenek parametresi kestirimleri veya alt testlere yönelik çalışmaların da oldukça az olduğu dikkat çekmektedir (Çakıcı Eser, 2015; Köse, 2010; Özkan, 2012). Çok boyutlu MTK'nın alt test puanı kestirimlerinde kullanılması konusunda daha fazla araştırmaya ihtiyaç olduğu düşünülmektedir. Alt testlerden oluşan ölçme araçlarından elde edilen puanların güvenirliği; ölçme aracının ölçtüğü teorik yapıların açıklanması ya da bileşke yapıya ilişkin kurulan modelin test edilmesi gibi yapı geçerliği; sınıflama geçerliği; çapraz geçerliği gibi ölçme araçlarının psikometrik özellikleri üzerinde MTK'ya dayalı yöntemlerin performansının daha fazla incelenmesi gerektiği düşünülmektedir. Belirlenen bu ihtiyaca bağlı olarak, ölçme araçlarının psikometrik özelliklerinden güvenirlik üzerinde çalışılmaya karar verilmiştir. Güvenilir alt test ve toplam test puanı kestirimleri konusuna katkı sağlamak amacıyla alt test ve toplam test arasındaki ilişki hiyerarşik MTK modelleri ile araştırılmak istenmiştir.

Araştırmada Kullanılan Hiyerarşik MTK Modelleri

Araştırmada kullanılan çok boyutlu madde tepki kuramı modelleri alt testler ile toplam test arasındaki ilişkiyi hiyerarşik düzeyde ele alması nedeniyle "hiyerarşik MTK modelleri" olarak adlandırılmıştır. Modellerin yapısal gösterimleri Şekil 1'de sunulmuştur.



Şekil 1. Araştırmada Kullanılan Hiyerarşik Çok Boyutlu MTK Modelleri

Araştırmada Hiyerarşik ÇBMTK Model için alt testlerin madde ve yetenek parametre kestirimlerinde matematiksel formülü Reckase (1997) tarafından ifade edilen çok boyutlu üç parametreli lojistik model kullanılmıştır. Hiyerarşik ÇBMTK Model'de toplam test puanı için yetenek kestirimleri Yao ve Schwarz (2006) tarafından tanımlanan Maksimum Bilgi Yöntemi ile elde edilmiştir. Maksimum test bilgi yöntemi ile mümkün olan tüm açı değerleri için elde edilen varyans değerini en küçük yapacak açı değeri hesaplanmaktadır. Böylece maksimum bilgiye sahip en güvenilir θ_α -birleşik (composite) puan- elde edilmektedir. Sonuç olarak bu yöntem ile Hiyerarşik ÇBMTK modelde toplam/genel test puanı ile alt test puanları arasında lineer bir ilişki kurulmaz. Aksine toplam puan ve alt test puanları arasındaki ilişkilerin farklı yetenek düzeylerinde veya farklı puan düzeylerinde farklılaşabileceği gerçeği dikkate alınarak toplam test puanları elde edilmektedir (Yao, 2010).

Üst Düzey Sıralı Model'de alt test puanları ile toplam test puanı arasında lineer bir ilişki kurulmaktadır. Bu ilişki toplam test puanı ile alt test puanları arasındaki korelasyonlara dayanmaktadır. Bu yaklaşıma göre alt testler kendi içlerinde tek boyutludur ama bütün alt testler dolaylı olarak genel bir boyutla ilişkilidir. Alan yazında böyle yapılara çoklu-tek boyutlu (multi-unidimensional) test yapıları da denilmektedir (Sheng ve Wickle, 2007).

İki Faktör Model'de bir maddenin hem spesifik bir alt boyutla hem de genel bir boyutla ilişkili olduğu varsayılır. Ölçme modeli açısından İki faktör Model hem Üst Düzey Sıralı Model hem de ÇBMTK model ile birbirine benzerlik gösterse de modeller arasında önemli farklılıklar vardır. ÇBMTK modelde alt testler arasındaki korelasyon model kestirimlerinde serbest bırakılırken İki faktör modelde genel boyut ile alt testlerin birbirine dik olduğu varsayılmaktadır. Üst Düzey Sıralı ve ÇBMTK modeldeki maddeler genel boyut ile doğrudan ilişkili olmazken İki Faktör modelde maddeler genel boyut ile doğrudan ilişkilidir. Ayrıca İki Faktör modelin temel amacı genel boyuta ilişkin yeteneği kestirmektir. Alt testler ikincil çıktılar olarak kabul edilmektedir ve genel yeteneğin artıklarından açıklanmaktadır. Schmid ve Leiman (1957) tarafından İki Faktör model ile Üst Düzey Sıralı modelin belirli matematiksel sınırlamalar altında eşit olduğu belirtilmiştir.

Araştırmanın Amacı

Ülkemizde yapılan geniş ölçekli sınavlarda olduğu gibi PISA, TIMSS, PIRLS gibi uluslararası geniş ölçekli sınavlarda da çoğu zaman politikacıların, eğitimcilerin veya velilerin daha çok ülke performansı/sıralaması üzerinde durduğu ve testlerin içeriğindeki alt ölçeklerin, alt yeterlik alanlarının ya da alt konuların üzerinde çok fazla durulmadığı dikkat çekmektedir. Oysaki eğitim araştırmacıları ve uygulayıcıları için böyle uluslararası geniş ölçekli testlerin anlamlı daha küçük alt birimlerinden yola çıkarak yapılacak çok boyutlu analiz sonuçlarının potansiyel olarak öğrencilerin öğrenme şeklinin ve dolaysıyla öğrenme çıktılarının belirlenmesi açısından daha anlamlı ve otantik olacağı söylenebilir. Bu bağlamda araştırmanın temel amacı alt testlerden elde edilebilecek tanısal bilgilere dikkat çekmek ve güvenilir alt test puanı kestirimlerine katkı sağlamaktır.

Alt testlerden oluşan geniş ölçekli testlere ilişkin hem genel hem de alt test puanı kestirimlerine ülkemizde kullanılan klasik yöntemlerden farklı bir bakış açısı getirmesi bu çalışmanın en önemli özelliğidir. Bu çalışmada, Türkiye'de uygulanan geniş ölçekli sınavların test yapısını çok boyutlu ve hiyerarşik modeller çerçevesinde ele almanın ve analiz etmenin alt test ve toplam test puan kestirimlerinin doğruluğuna ve güvenirliğine etkisinin nasıl olacağı da araştırılmak istenmiştir. TEOG veri setinin özelliklerine göre güvenirliğe etki edebilecek hiyerarşik çok boyutlu MTK Model, İki faktör Model ve Üst Düzey Sıralı Model'in alt test ve toplam test puanı kestirimleri üzerindeki performansı incelenmiştir. Bu amaçla "alt test uzunluğu (20,30,40), alt test sayısı (2,3) ve alt testler arasındaki korelasyonların büyüklüğü (0.0, 0.3, 0.5, 0.8) koşulları altında üretilen veriler ile gerçek verilerin toplam test puanı ve alt test puanları kestirimlerinin doğruluğu ve güvenirlikleri Üst Düzey Sıralı Model, İki Faktör Model ve çok boyutlu hiyerarşik MTK modele göre nasıl değişmektedir?" sorusuna yanıt aranmıştır.

YÖNTEM

Deneysel desenler değişkenler arasındaki neden sonuç ilişkilerini test etmeyi amaçlayan araştırma desenleridir. Bu amacı gerçekleştirebilmek için Fraenkel, Wallen ve Hyun'a (2011, s.265-266) göre deneysel desenler, bağımsız değişken/lerin bağımlı değişken/ler üzerindeki etkisini incelemek için en az iki koşulun karşılaştırılmasını ve bağımsız değişkenin araştırmacı tarafından doğrudan değişimlenmesini (manipüle edilmesini) gerektirir. Ayrıca, araştırmacılar iç geçerliği korumak için dışsal değişkeni (ilgilenilmeyen ya da istenmeyen değişken) kontrol altına alarak bağımlı değişken üzerinde ölçme yapmalıdır (Kerlinger, 1973, s.300-313; Gall, Gall ve Borg, 2003, s.367-368). Simülasyon çalışmaları doğası gereği araştırmacılara bağımsız değişkenleri değişimleme ve dışsal değişkenleri kontrol altına alma imkânı sağlar. Bir bölümünde simülasyon verisi kullanılan bu araştırmada farklı test koşullarında üretilmiş verilerin toplam test puanı ve alt test puanları kestirimlerinin doğruluğu ve güvenirliği farklı modeller ve farklı test koşulları açısından karşılaştırıldığından çalışma simülatif verilerle yürütülen deneysel araştırma özelliği taşımaktadır. Ayrıca araştırmanın gerçek veri uygulaması içeren diğer bölümü, TEOG sınavı ile ilgili mevcut duruma ait bilgiler vereceğinden bu çalışmanın betimsel araştırma özelliği de bulunmaktadır.

Simülasyon Koşulları

Çalışma Grubu

Alan yazın çalışmalarından yetenek parametresi kestirimlerinde 1000 ve üzeri örneklemler arasında fark gözlenmediği (de a Torre ve Song, 2009; Yao ve Boughton, 2009) bulgusuna dayalı olarak bu araştırmada örneklem büyüklüğü bağımsız değişken olarak seçilmemiş ve araştırmanın verileri N=3000 olacak şekilde üretilmiştir.

Alt Test Sayısı

Alan yazındaki çok boyutlu ya da alt testlerden oluşan ölçme yapıları için parametre doğrulama çalışmaları incelendiğinde hem simülasyon hem de gerçek veri setleri üzerinde yapılan araştırmalarda kullanılan ölçme araçlarının iki ila altı boyut arasında değişen alt boyutlarda/alt testlerde olduğu görülmüştür (Edwards ve Vevea, 2006; Lee, 2012; Yao, 2017).Ülkemizde yapılan çeşitli geniş ölçekli sınavlar incelendiğinde TEOG sınavının aynı oturumda yapılan sınavlarının üç toplamda altı, KPSS'de aynı oturumda yapılan çeşitli sınavların 2 (Genel Kültür ve Genel yetenek gibi) ya da beş (Çalışma ekonomisi, Ekonometri, İstatistik, Kamu Yönetimi ve Uluslararası ilişkiler gibi), ALES sınavının ise 2 (sayısal ve sözel) alt testten oluştuğu görülmüştür. Alan yazında yapılan araştırmalar ve ülkemizde yapılan geniş ölçekli sınavlar ve test uzunlukları göz önüne alındığında gerçek durumları temsil etmesi açısından bu araştırmada alt test sayısı iki ve üç olarak belirlenmiştir.

Alt Test Uzunluğu

Alan yazındaki çok boyutlu ya da alt testlerden oluşan ölçme yapıları için parametre doğrulama çalışmaları incelendiğinde hem simülasyon hem de gerçek veri setleri üzerinde yapılan araştırmalarda kullanılan ölçme araçlarının 5-60 madde arası alt test uzunluğuna sahip olduğu görülmektedir. Ülkemizde yapılan geniş ölçekli sınavlar incelendiğinde TEOG sınavlarının toplam altı alt testten ve 20'şer maddeden oluştuğu, ALES sınavının sayısal bölümünün 40'ar sorudan oluşan sayısal1 ve sayısal2 olarak iki at testten, sözel bölümünün ise 40'ar sorudan oluşan sözel1 ve sözel2 olarak iki alt testten oluştuğu belirlenmiştir. KPSS sınavları incelendiğinde aynı oturumda yapılan sınavlardan genel kültür ve genel yetenek alt testlerinin 60'ar sorudan, A grubu sınavlarından aynı oturumda yapılan sınavlardan Hukuk, İktisat, İşletme, Maliye ve Muhasebe alt testlerinin 30'ar sorudan ve Din Hizmetleri Alan Bilgisi Testi'nin (DHAB) ise DHAB1 ve DHAB2 alt testlerinin 20'şer sorudan oluştuğu görülmüştür. Alan yazın ve ülkemizde yapılan sınavlara dayalı olarak gerçek durumları temsil etmesi açısından bu araştırmada alt test uzunluğu 20, 30 ve 40 madde olarak belirlenmiştir.

Alt Testler Arasındaki Korelasyon

Alan yazındaki çok boyutlu ya da alt testlerden oluşan ölçme yapıları için parametre doğrulama çalışmaları incelendiğinde alt testler arasındaki korelasyonların parametre kestirimi üzerinde etkisinin olduğu belirtilmiştir (de la Torre ve Patz, 2005; Shin, 2007; Shin, Ansley, Tsai, ve Mao, 2005; Yao, 2010). Yine alan yazındaki araştırmalarda alt testler ya da alt testler arası korelasyon koşulu için 0.0-1.0 arasında değişen düzeylerde büyüklükler seçildiği görülmüştür. Ülkemizde 29 Nisan 2015'te yapılan TEOG sınavının altı alt testi arasındaki korelasyonların 0.0 civarında olması da göz önünde bulundurularak bu araştırma için alt testler arası korelasyon düzeyleri 0.0, 0.3, 0.5 ve 0.8 olarak belirlenmiştir.

Tablo 1: Simülasyon Koşulları ve Düzeyleri

Alt Test Sayısı	Al Test Uzunluğu	Alt Testler Arası Korelasyon
2	20	0.0
3	30	0.3
	40	0.5
		0.8

Tekrar (replikasyon) sayısı:50

Verilerin Üretilmesi

Araştırmada kullanılan veri setleri 29 Nisan 2015'te yapılan TEOG sınavının psikometrik özelliklerine dayalı olarak üretilmiştir. Bu sınav verisinin faktör yapısını belirlemek için Factor 10.3

(Lorenzo-Seva ve Ferrando, 2006) programında Ağırlıklandırılmamış En Küçük Kareler (ULS) yöntemi ve varimax döndürme tekniğine göre veriler analiz edilmiştir. Analiz sonucunda elde edilen faktör yükü ve alt testler arası korelasyon matrislerinin incelenmesi sonucunda TEOG verisinin basit yapılı örtük yetenek konfigürasyonuna sahip olduğu görüldüğünden bu çalışmada veriler basit yapılı olacak şekilde üretilmiştir. TEOG 2015 sınavının 3PL modele göre elde edilen madde parametreleri ve alan yazında yer alan simülasyon çalışmaları göz önünde bulundurularak bu çalışmada kullanılan verilerin madde ayırt edicilik parametresi ranjı [0.8-3] arasında olacak şekilde ortalaması 1.5 ve varyansı 0.5 olan bir normal dağılımdan; güçlük parametresi ranjı [-2-2] arasında olacak şekilde ortalaması 0.0 ve varyansı 1.0 olan bir normal dağılımdan ve en düşük asimtot (şans) parametresi ise (6,16) olan bir beta dağılımdan üretilmiştir. Yetenek parametreleri çok değişkenli normal dağılıma θi ~ MVN (0, Σ) dayalı olarak ortalaması sıfır (0), varyansı ise araştırma koşullarında belirlenmiş olan varyans-kovaryans matrisine göre üretilmiştir. Üretilen madde ve yetenek parametreleri kullanılarak Tablo 1'de özetlenen koşullar altında iki kategorili 3000 kişilik veri setleri 50 tekrara dayalı olarak SimuMIRT (Yao, 2003) programı kullanılarak üretilmiştir. Harwell, Stone, Hsu ve Kirisci (1996) monte carlo simülasyon çalışmaları için optimal koşulları belirleme, mevcut programları inceleme ve simülasyon çalışmalarının kavramsallaştırılmasının önemini açıklama konusundaki çalışmalarında, simülasyon çalışmalarında en az 25 replikasyon kullanılması gerektiğini belirtmişlerdir. Bu konuda alan yazın incelendiğinde, Yao'nun (2010) ve Huang, Wang ve Chen (2013) 20 tekrar, Çakıcı Eser'in (2015) 25 tekrar ve de la Torre'nin (2009) 100 tekrar ile çalışmalarını yürüttükleri görülmüştür. Bu araştırmada ise tekrar sayısı 50 olarak belirlenmiştir.

Verilerin Analizi

Gerçek parametreler ile kestirilen parametrelerin karşılaştırılabilmesi için kestirilen parametreler ile gerçek değerlerin aynı ölçekte olması gerekir. Bunu sağlayabilmek için parametre kestirimlerinde popülasyon parametrelerinin onların gerçek değerine sabitlenmesi gerekir. Normalde gerçek değerleri bilinemez ama üretilen verilerin ortalama ve varyans-kovaryans matrisi, madde parametrelerinin dağılımları gerçek değer yerine kullanılabilir. Yao (2010) yetenek parametreleri kestirimlerinde madde parametrelerinin sabitlenmesi ile sabitlenmemesi yaklaşımı arasında bir fark olmadığını belirtse de bu araştırmada üretilen verilerin özellikleri önsel bilgi (prior) olarak kullanılmıştır. Ayrıca kestirim modelleri ve simülasyon koşullarına göre elde edilen RMSE değerleri ortalaması arasında anlamlı fark olup olmadığı varyans analizi ile test edilmiştir. Varyans analizi sonucunda en az iki grup arasında anlamlı farklılığın gözlendiği durumlar için farklılığın hangi gruplar arasında olduğunu belirlemek amacıyla çoklu karşılaştırma testi yapılmıştır. Karşılaştırma testi olarak Fisher'in LSD Testi kullanılmıştır. SPSS programı ara yüzü seçeneklerine göre yalnızca ana etkiler için karşılaştırma testi yapabilmesi nedeniyle etkileşimler için karşılaştırma testleri syntax yazılarak yapılmıştır.

Değerlendirme Kriteri

Araştırmada alt test puanlarının ve toplam test puan kestirimlerinin doğruluklarını değerlendirmek için RMSE (Ortalama hata kareler kökü) ve güvenirlik istatistikleri kullanılmıştır. RMSE değeri, gerçek parametre ile kestirilen parametreler arasındaki farkların ortalamasının karekökünü ifade etmektedir. Güvenirlik değeri ise gerçek parametre ile kestirilen parametreler arasındaki korelasyonun kareler ortalamasını ifade etmektedir. Bu istatistiklere ait matematiksel ifadeler aşağıdaki gibidir:

$$\text{RMSE}\left(\tau_j\right) = \sqrt{\frac{1}{n*N}\sum_{d=1}^{n}(\tau_j^* - \tau_j)^2} \qquad Güvenirlik = \frac{1}{n}\sum_{d=1}^{n} cor\,(\tau_j^*, \tau_j)^2$$

τ_j: j parametresinin gerçek değeri

τ_j^*: j parametresinin kestirilen değeri

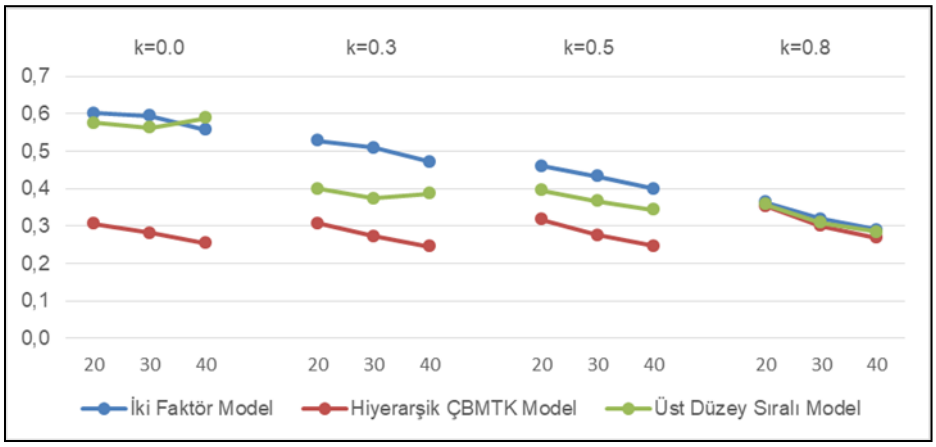
n: tekrar (replikasyon) sayısı

N: örneklem büyüklüğü

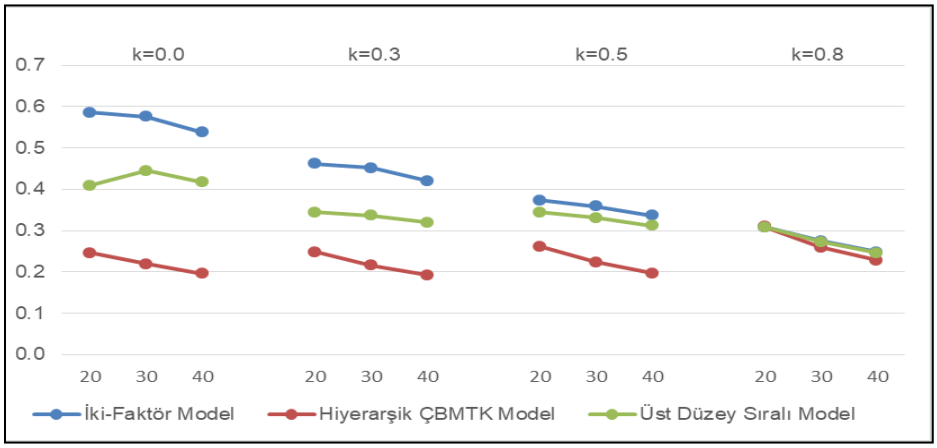
BULGULAR

Toplam Test Puanı İçin Kestirilen Yetenek Parametrelerine Ait RMSE Değerlerine Yönelik Bulgular

Şekil 2 ve Şekil 3'te araştırmada ele alınan koşullara dayalı olarak sırasıyla iki ve üç boyutlu veri setlerinden toplam test puanı için üç hiyerarşik madde tepki kuramı modeli kullanılarak kestirilen yetenek parametrelerine ilişkin RMSE değerleri verilmiştir.



Şekil 2. İki Boyutlu Veri Setlerinden Toplam Test Puanı İçin Kestirilen Yetenek Parametrelerine Ait RMSE Değerleri



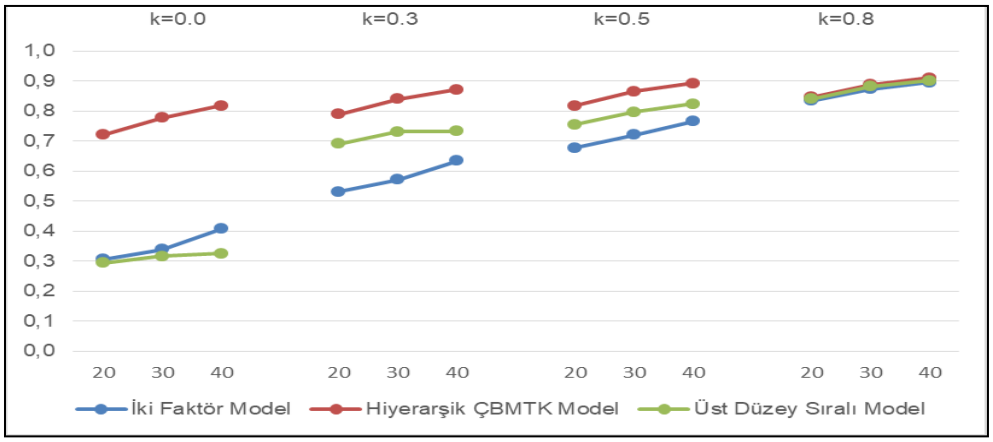
Şekil 3. Üç Boyutlu Veri Setlerinden Toplam Test Puanı İçin Kestirilen Yetenek Parametrelerine Ait RMSE Değerleri

Şekil 2 ve Şekil 3'e göre üç kestirim modelinin tüm koşullar altındaki performansları karşılaştırıldığında, en düşük hata düzeyine sahip kestirimlerin İki Faktör modelinden elde edildiği görülürken en düşük hata düzeyine sahip kestirimlerin Hiyerarşik ÇBMTK modelden edildiği görülmektedir. Alt testler arası korelasyon düzeyi attıkça Hiyerarşik ÇBMTK dışında modellerin kestirim hatalarının genel olarak azaldığı ve üç modele ait sonuçların birbirine yaklaştığı gözlenmektedir. Özellikle alt testler arası korelasyon düzeyinin 0.8 olduğu durumda üç yöntemin benzer hatalar ile parametre kestirimi yaptığı söylenebilir. Alt test uzunluğundaki artışın üç yöntem için de genel olarak parametre kestirim hatasını azalttığı görülmektedir. İki-faktör Model'e ait kestirim hataları ile alt test uzunluğu ve alt testler arası korelasyon arasında azalan doğrusal yönde bir ilişki olduğu gözlenirken Üst Düzey Sıralı Model için bazı koşullarda değişken düzeyde ama

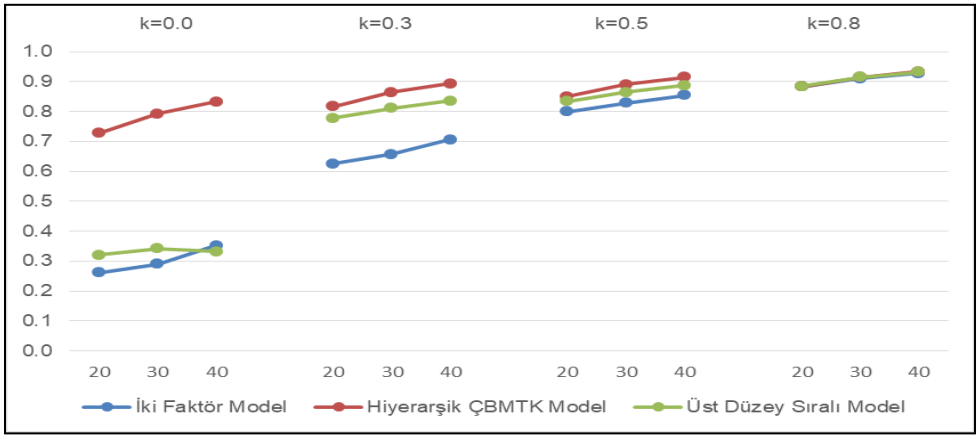
çoğu koşul için benzer bir ilişki gözlenmektedir. Hiyerarşik ÇBMTK Model'de ise kestirim hataları ile alt test uzunluğu arasında azalan fakat alt testler arası korelasyon arasında artan doğrusal yönde bir ilişki olduğu görülmektedir. İki ve üç boyutlu veri setlerine ait değerler birlikte değerlendirildiğinde alt test sayısındaki artışın üç modelin toplam test puanı için yetenek parametresi kestirim hatalarını azalttığı gözlenmektedir. İki faktör modeli için alt test sayısı ikiden üçe çıkarıldığında yetenek parametresi kestirim hatası alt testler arası korelasyon koşulunun her bir düzeyi için sırasıyla ortalama %3, %12, %17 ve %15 oranında azalmaya neden olurken bu durum Hiyerarşik ÇBMTK model için sırasıyla ortalama %22, %21, %19 ve %14 oranında, Üst Düzey Sıralı Model için sırasıyla %26, %14, %11 ve %13 oranında azalmaya neden olduğu görülmektedir.

Toplam Test Puanı İçin Kestirilen Yetenek Parametrelerine Ait Güvenirlik Değerlerine Yönelik Bulgular

Şekil 4 ve Şekil 5'te araştırmada ele alınan koşullara dayalı olarak sırasıyla iki ve üç boyutlu veri setlerinden toplam test puanı için üç hiyerarşik madde tepki kuramı modeli kullanılarak kestirilen yetenek parametrelerine ilişkin güvenirlik değerleri verilmiştir.



Şekil 4. İki Boyutlu Veri Setlerinden Toplam Test Puanı İçin Kestirilen Yetenek Parametrelerine Ait Güvenirlik Değerleri

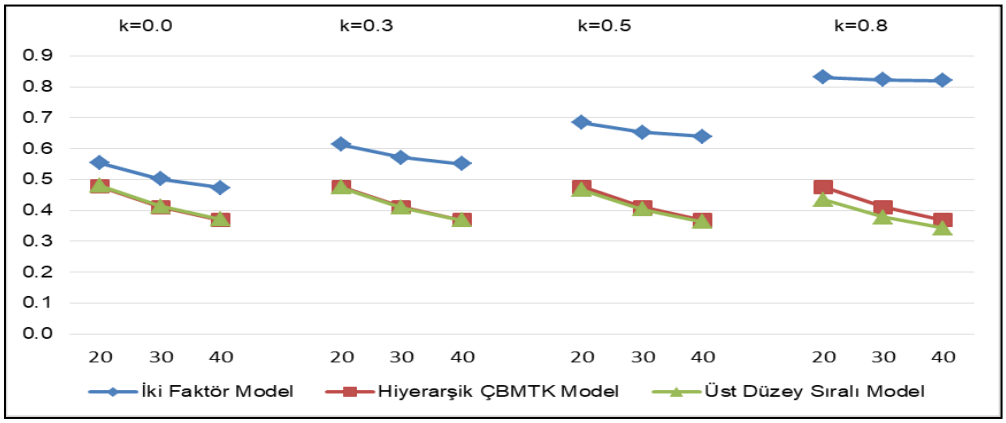


Şekil 5. Üç Boyutlu Veri Setlerinden Toplam Test Puanı İçin Kestirilen Yetenek Parametrelerine Ait Güvenirlik Değerleri

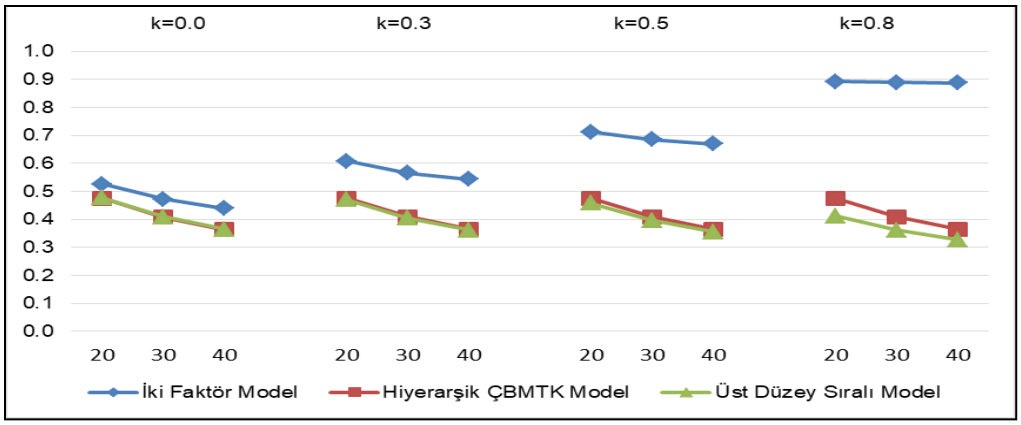
Şekil 4 ve Şekil 5'e göre hem iki hem de üç boyutlu veri setleri için üç kestirim modelinin tüm koşullar altındaki performansları karşılaştırıldığında kabul edilebilir en güvenilir kestirimlerin Hiyerarşik ÇBMTK modelden edildiği görülmektedir. Alt testler arası korelasyon düzeyi attıkça modellerin yetenek parametresi kestirim güvenirliğinin arttığı, Üst Düzey Sıralı Model kestirim güvenirliğinin korelasyonun 0.3 ve üzeri düzeylerde kabul edilebilir düzeylerde olduğu ve üç modelin kestirim güvenirliğinin korelasyonun 0.5 ve üzeri düzeylerde birbirine yaklaştığı gözlenmektedir. Özellikle alt testler arası korelasyon düzeyinin 0.8 olduğu durumda üç yöntemin benzer güvenirlik ile parametre kestirimi yaptığı söylenebilir. Alt test uzunluğu ve alt testler arası korelasyon düzeyindeki artışın her üç yöntem için de genel olarak parametre kestirim güvenirliğini arttırdığı görülmektedir. İki-faktör Model ve Üst Düzey Sıralı Model'in alt testler arası korelasyonun 0.0 düzeyinde yetenek parametre kestirimlerinin kabul edilemez düzeyde güvenirlik ile kestirildiği dikkat çekmektedir. Her üç modele ait kestirim güvenirliği ile alt test uzunluğu ve alt testler arası korelasyon arasında artan doğrusal yönde bir ilişki olduğu gözlenmektedir. İki ve üç boyutlu veri setlerine ait değerler birlikte değerlendirildiğinde boyut sayısındaki artışın üç modelin toplam test puanı için yetenek parametresi kestirim güvenirliğini koşulların çoğunda arttırdığı gözlenmektedir. İki faktör modelde alt test sayısındaki artışın kabul edilebilir düzeyde parametre kestirimi koşullarında iyileşme sağladığı görülmektedir. İki Faktör Model için alt test sayısını ikiden üçe çıkarmanın yetenek parametresi kestirim güvenirliğinin alt testler arası korelasyon düzeyinin 0.0 olduğu durum dışındaki diğer düzeyleri için sırasıyla ortalama %15, %15 ve %5 oranında artışa neden olduğu görülmektedir. Bu durum Hiyerarşik ÇBMTK model için alt testler arası korelasyon koşulunun tüm düzeylerinde ortalama %3 oranında artışa neden olurken Üst Düzey Sıralı Model için boyut sayısı artışı ile alt testler arası korelasyon koşulunun tüm düzeyleri için yetenek parametresi kestirim güvenirliğinde sırasıyla ortalama %7, %13, %9 ve %4 oranında artış olduğu gözlenmektedir.

Alt Test Puanı İçin Kestirilen Yetenek Parametrelerine Ait RMSE Değerlerine Yönelik Bulgular

Şekil 6 ve Şekil 7'de araştırmada ele alınan koşullara dayalı olarak sırasıyla iki ve üç boyutlu veri setlerinden alt test puanı için üç hiyerarşik madde tepki kuramı modeli kullanılarak kestirilen yetenek parametrelerine ilişkin RMSE değerleri verilmiştir.



Şekil 6. İki Boyutlu Veri Setlerinden Alt Test Puanı İçin Kestirilen Yetenek Parametrelerine Ait RMSE Değerleri



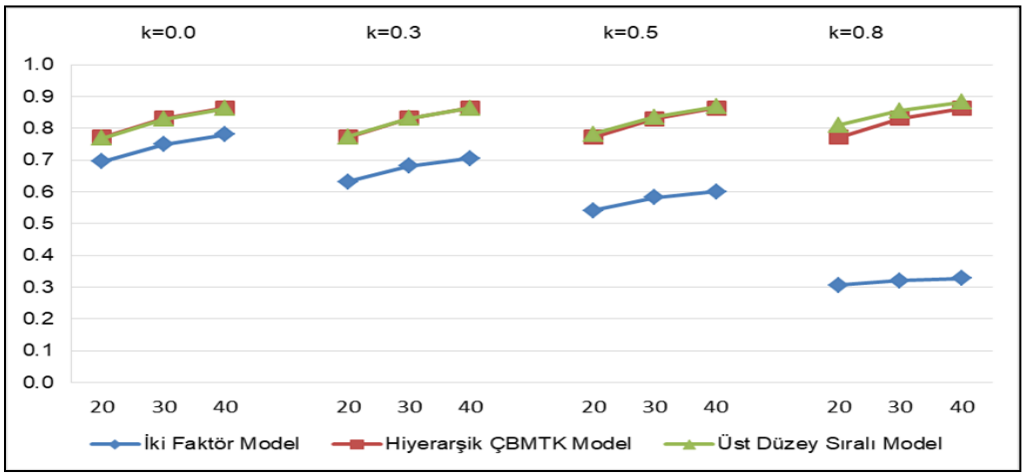
Şekil 7. Üç Boyutlu Veri Setlerinden Alt Test Puanı İçin Kestirilen Yetenek Parametrelerine Ait RMSE Değerleri

Şekil 6 ve Şekil 7'ye göre üç kestirim modelinin tüm koşullar altındaki performansları karşılaştırıldığında, en düşük hata düzeyine sahip kestirimlerin İki Faktör modelinden elde edildiği görülürken en düşük hata düzeyine sahip kestirimlerin Üst Düzey Sıralı modelden edildiği görülmektedir. Alt testler arası korelasyon düzeyi attıkça Üst Düzey Sıralı Model ve Hiyerarşik ÇBMTK Model için alt test yetenek kestirim hatalarının azaldığı gözlenirken İki-faktör Model için hataların arttığı gözlenmektedir. Alt testler arası korelasyon koşulunun ilk iki düzeyinde Üst Düzey Sıralı Model ve Hiyerarşik ÇBMTK Model'in alt test yetenek kestirim hatalarının benzer olduğu görülürken korelasyon düzeyi arttıkça Üst Düzey Sıralı Model'in daha iyi performans gösterdiği görülmektedir. İki-faktör Model'e ait kestirim hataları ile alt test uzunluğu arasında azalan fakat alt testler arası korelasyon arasında artan doğrusal yönde bir ilişki olduğu gözlenirken Üst Düzey Sıralı Model'e ait kestirim hataları ile alt test uzunluğu ve alt testler arası korelasyon koşulları arasında azalan doğrusal yönde bir ilişki olduğu gözlenmektedir. Hiyerarşik ÇBMTK Model'e ait kestirim hataları ile korelasyon koşulu arasında bir ilişki olmadığı ama alt test uzunluğu ile kestirim hataları arasında azalan doğrusal yönde bir ilişki olduğu görülmektedir.

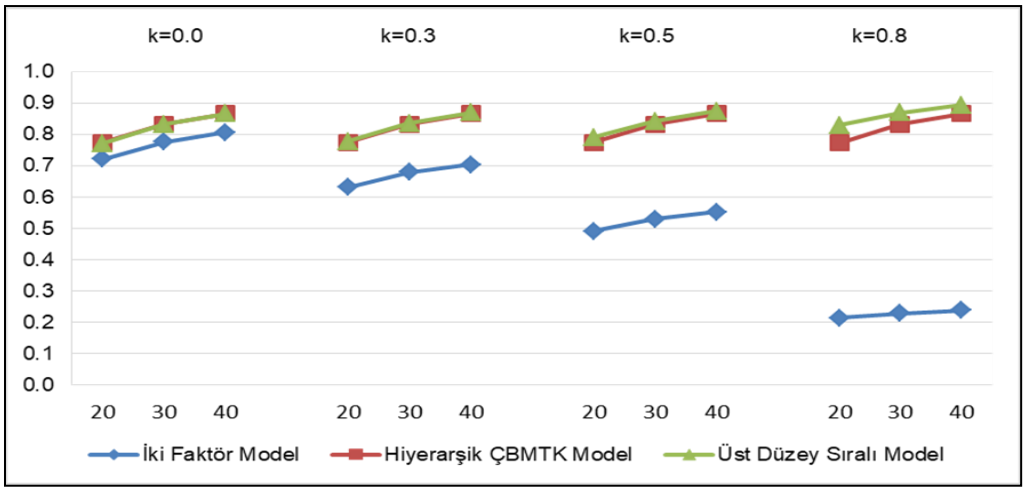
İki ve üç boyutlu veri setlerine ait değerler birlikte değerlendirildiğinde boyut sayısındaki artışın üç modelin alt test puanı için yetenek parametresi kestirim hataları üzerindeki etkisinin değişkenlik gösterdiği görülmektedir. İki-faktör modeli için boyut sayısı ikiden üçe çıkarıldığında yetenek parametresi kestirim hatası alt testler arası korelasyon koşulunun düşük düzeylerinde azalmaya neden olurken korelasyonun yüksek düzeylerinde artmaya neden olduğu görülmektedir. Bu model için boyut sayısı arttırıldığında alt testler arası korelasyonun 0.0 düzeyinde 20, 30 ve 40 maddelik alt testlerden elde edilen yetenek parametre kestirim hatalarının sırasıyla yaklaşık olarak %5, %6 ve %7 oranlarında azalmasına neden olurken korelasyonun 0.3 düzeyinde yaklaşık olarak %1 oranda azalmasına neden olmaktadır. Yine İki faktör model için boyut sayısı arttırıldığında alt testler arası korelasyonun 0.5 düzeyinde 20, 30 ve 40 maddelik alt testlerden elde edilen yetenek parametre kestirim hatalarının sırasıyla yaklaşık olarak %4, %5 ve %5 oranlarında artmasına neden olurken korelasyonun 0.8 düzeyinde sırasıyla yaklaşık olarak %1 oranda artmasına neden olmaktadır. Hiyerarşik ÇBMTK model için boyut sayısının ikiden üçe çıkarılmasının alt test yetenek kestirim hatalarında bir etkisi olmadığı görülmektedir. Üst Düzey Sıralı Model için boyut sayısı ikiden üçe çıkarıldığında alt testler arası korelasyon koşulunun düşük düzeylerinde minimal düzeyde olmakla birlikte alt test yetenek parametresi kestirim hatasının korelasyonun artışı ile azalmaya neden olduğu görülmektedir. Bu model için boyut sayısı arttırıldığında alt testler arası korelasyonun 0.0 ve 0.3 düzeylerinde 20, 30 ve 40 maddelik alt testlerden elde edilen yetenek parametre kestirim hatalarının yaklaşık olarak %0-%2 arası oranlarında azalmasına neden olurken korelasyonun 0.3 düzeyinde yaklaşık olarak %1-%2 arası oranlarda ve 0.8 düzeyinde yaklaşık olarak %4-%5 arası oranda azalmasına neden olmaktadır.

Alt Test Puanı İçin Kestirilen Yetenek Parametrelerine Ait Güvenirlik Değerlerine Yönelik Bulgular

Şekil 8 ve Şekil 9'da araştırmada ele alınan koşullara dayalı olarak sırasıyla iki ve üç boyutlu veri setlerinden alt test puanı için üç hiyerarşik madde tepki kuramı modeli kullanılarak kestirilen yetenek parametrelerine ilişkin güvenirlik değerleri verilmiştir.



Şekil 8. İki Boyutlu Veri Setlerinden Toplam Test Puanı İçin Kestirilen Yetenek Parametrelerine Ait Güvenirlik Değerleri



Şekil 9. Üç Boyutlu Veri Setlerinden Toplam Test Puanı İçin Kestirilen Yetenek Parametrelerine Ait Güvenirlik Değerleri

Şekil 8 ve Şekil 9'a göre üç kestirim modelinin tüm koşullar altındaki performansları karşılaştırıldığında kabul edilebilir en güvenilir kestirimlerin Üst Düzey Sıralı Modelden elde edildiği görülmektedir. Alt testler arası korelasyon düzeyi attıkça Üst Düzey Sıralı Model ve Hiyerarşik ÇBMTK Model'in yetenek parametresi kestirim güvenirliğinin arttığı ve kestirim güvenirliğinin korelasyonun tüm düzeylerinde kabul edilebilir düzeylerde olduğu gözlenmektedir. Ayrıca bu iki modelin alt test yetenek kestirim güvenirliğinin korelasyonun ilk üç düzeyinde aynı/benzer olduğu ve korelasyonun 0.8 düzeyinde Üst Düzey Sıralı Modelin daha güvenilir sonuçlar verdiği söylenebilir. İki faktör Modelin alt test yetenek kestirim güvenirliğinin korelasyon

düzeyi arttıkça önemli düzeyde azaldığı ve bu yöntemin kabul edilebilir düzeyde güvenilir kestirimlerinin düşük korelasyon düzeyinde elde edildiği gözlenmektedir. Fakat düşük korelasyon düzeyinde dahi İki faktör Model parametre kestirim güvenirliğinin Üst Düzey Sıralı Model ve Hiyerarşik ÇBMTK Model kestirimlerine göre daha düşük olduğu dikkat çekmektedir. Alt test uzunluğu ve alt testler arası korelasyon düzeyindeki artışın her üç yöntem için de parametre kestirim güvenirliğini arttırdığı görülmektedir. Alt test yetenek parametresi kestirim güvenirliği ile alt test uzunluğu ve alt testler korelasyon arasında Üst Düzey Sıralı Model ve Hiyerarşik ÇBMTK Model için artan doğrusal yönde bir ilişki olduğu gözlenirken İki faktör Model için güvenirlik ile alt test uzunluğu arasında artan fakat korelasyon ile azalan doğrusal yönde bir ilişki olduğu gözlenmektedir.

İki ve üç boyutlu veri setlerine ait değerler birlikte değerlendirildiğinde boyut sayısındaki artışın üç modelin alt test puanı için yetenek parametresi kestirim güvenirliğine etkisinin değişken olduğu gözlenmektedir. İki faktör modeli için boyut sayısını ikiden üçe çıkarmanın alt test kestirim güvenirliğini korelasyon düzeyinin 0.0 olduğu durumda yaklaşık olarak %4 oranında arttırdığı, korelasyonun 0.3 düzeyinde etkisinin olmadığı, korelasyonun 0.5 düzeyinde yaklaşık olarak %9 oranında azalttığı ve korelasyonun 0.8 düzeyinde yaklaşık olarak %28 oranında azalttığı görülmektedir. Hiyerarşik ÇBMTK model için boyut sayısının ikiden üçe çıkarılmasının alt test yetenek kestirim hatalarında bir etkisi olmadığı görülmektedir. Üst Düzey Sıralı Model için boyut sayısını ikiden üçe çıkarmanın alt test kestirim güvenirliğine korelasyon düzeyinin 0.0 düzeyinde etkisinin olmadığı, korelasyonun 0.3 ve 0.5 düzeyinde güvenirliği yaklaşık olarak %1 oranında arttırdığı, korelasyonun 0.8 düzeyinde güvenirliği yaklaşık olarak %2 oranında artırdığı görülmektedir.

Toplam Test Puanlarına ait RMSE ve Güvenirlik Değerleri İçin Varyans Analizi Sonuçları

Simülasyon çalışmasında tüm koşullar altında elde edilen toplam test puanlarına ait RMSE ve güvenirlik değerleri üzerinde model, alt test sayısı, alt test uzunluğu ve alt testler arasındaki korelasyonun etkisini incelemek için yapılan varyans analizi sonuçları sırasıyla Tablo 2 ve Tablo 3'te verilmiştir.

Tablo 2. Toplam Test Puanlarına Ait RMSE Değerleri İçin Varyans Analizi Sonuçları

Varyans kaynağı	Kareler toplamı	df	Kareler ortalaması	F	p	Kısmi η^2
Alt test sayısı	2,964	1	2,964	2239,519	0,000	0,388
Alt test uzunluğu	1,482	2	0,741	559,900	0,000	0,241
Korelasyon	10,010	3	3,337	2521,279	0,000	0,682
Model	19,153	2	9,576	7236,319	0,000	0,804
Alt test sayısı * alt test uzunluğu	0,016	2	0,008	6,104	0,002	0,003
Alt test sayısı * korelasyon	0,127	3	0,042	31,884	0,000	0,026
Alt test sayısı * model	0,066	2	0,033	24,815	0,000	0,014
Alt test uzunluğu * korelasyon	0,167	6	0,028	21,091	0,000	0,035
Alt test uzunluğu * model	0,136	4	0,034	25,756	0,000	0,028
Korelasyon * model	8,567	6	1,428	1078,886	0,000	0,647
Alt test sayısı * Alt test uzunluğu * korelasyon	0,009	6	0,002	1,196	0,305	0,002
Alt test sayısı * Alt test uzunluğu * model	0,018	4	0,004	3,358	0,009	0,004
Alt test sayısı * korelasyon * model	0,687	6	0,114	86,494	0,000	0,128
Alt test uzunluğu * korelasyon * model	0,056	12	0,005	3,501	0,000	0,012
Alt test sayısı * Alt test uzunluğu * korelasyon * model	0,016	12	0,001	0,999	0,447	0,003

Tablo 2 incelendiğinde toplam test puanlarına ait RMSE değerleri üzerinde tüm ana ve tüm ikili ortak etkilerin anlamlı düzeyde etkisi olduğu gözlenmektedir. Ana etkiler açısından etki büyüklükleri incelediğinde en fazla etkiye sahip değişkenlerin sırasıyla model (kısmi η^2 =0.804) ve alt testler arası korelasyon (kısmi η^2 =0.682) olduğu görülmektedir. İkili ortak etkiler incelendiğinde toplam testlere ait RMSE değerlerinin varyansını en fazla açıklayan etkileşimin korelasyon*model (kısmi η^2 =0.647) olduğu ve diğer ikili etkileşimlerin etki büyüklerinin (kısmi η^2 ≤0.035) çok düşük olduğu gözlenmektedir. Üçlü ortak etkiler içerisinde en fazla etkiye sahip etkileşimin alt test uzunluğu*korelasyon*model (kısmi η^2 =0.128) olduğu görülürken diğer üçlü etkileşimlerin etkisinin ya olmadığı (alt test sayısı*alt test uzunluğu*korelasyon etkileşimi, p=0.305) ya da çok düşük (kısmi η^2 ≤0.012) olduğu görülmektedir. Dörtlü ortak etkinin RMSE değerlerinin varyansına anlamlı bir katkısı olmadığı görülmektedir (p=0.447).

Çoklu karşılaştırma testi sonucunda ana etkiler için koşulların tüm düzeyleri arasında anlamlı farklılık olduğu bulunmuştur. Korelasyon, alt test uzunluğu ve alt test sayısı koşullarının düzeyleri arttıkça hatanın azaldığı gözlenmiştir. Modeller açısından en az hatalı kestirim yapan modeller sırasıyla Hiyerarşik ÇBMTK, Üst Düzey Sıralı ve İki Faktör modeldir. Korelasyon*model ikili etkileşimin modellere göre ikili karşılaştırma testi sonucunda yalnızca 0.8 korelasyon düzeyi için İki Faktör Model ile Üst Düzey Sıralı Model arasında anlamlı farklılık gözlenmezken diğer tüm ikili karşılaştırmalar anlamlı bulunmuştur. Yine, korelasyon*model etkileşimin korelasyona göre ikili karşılaştırma testi sonucunda ise İki Faktör ve Üst Düzey Sıralı modeller için korelasyonun tüm ikili etkileşimleri arasında anlamlı farklılık gözlenirken Hiyerarşik ÇBMTK model için yalnızca 0.8 korelasyon ile diğer korelasyon düzeyleri ve 0.5 korelasyon ile 0.3 korelasyon düzeyi arasında anlamlı farklılık gözlenmiştir.

Tablo 3. Toplam Test Puanlarına Ait Güvenirlik Değerleri İçin Varyans Analizi Sonuçları

Varyans kaynağı	Kareler toplamı	df	Kareler ortalaması	F	p	Kısmi η^2
Alt test sayısı	1,509	1	1,509	497,963	0,000	0,124
Alt test uzunluğu	2,721	2	1,361	448,974	0,000	0,203
Korelasyon	90,152	3	30,051	9915,606	0,000	0,894
Model	24,928	2	12,464	4112,629	0,000	0,700
Alt test sayısı * alt test uzunluğu	0,020	2	0,010	3,371	0,034	0,002
Alt test sayısı * korelasyon	0,806	3	0,269	88,663	0,000	0,070
Alt test sayısı * model	0,154	2	0,077	25,379	0,000	0,014
Alt test uzunluğu * korelasyon	0,038	6	0,006	2,117	0,048	0,004
Alt test uzunluğu * model	0,162	4	0,040	13,349	0,000	0,015
Korelasyon * model	25,631	6	4,272	1409,532	0,000	0,706
Alt test sayısı * Alt test uzunluğu * korelasyon	0,009	6	0,002	0,499	0,810	0,001
Alt test sayısı * Alt test uzunluğu * model	0,008	4	0,002	0,637	0,636	0,001
Alt test sayısı * korelasyon * model	0,497	6	0,083	27,322	0,000	0,044
Alt test uzunluğu * korelasyon * model	0,140	12	0,012	3,842	0,000	0,013
Alt test sayısı * Alt test uzunluğu * korelasyon * model	0,015	12	0,001	0,400	0,964	0,001

Tablo 3 incelendiğinde toplam test puanlarına ait güvenirlik değerleri üzerinde tüm ana ve tüm ikili ortak etkilerin anlamlı düzeyde etkisi olduğu gözlenmektedir. Ana etkiler açısından etki büyüklükleri incelediğinde en fazla etkiye sahip değişkenlerin sırasıyla alt testler arası korelasyon (kısmi η^2

=0.894) ve model (kısmi η^2=0.700) olduğu görülmektedir. İkili ortak etkiler incelendiğinde toplam testlere ait güvenirlik değerlerinin varyansını en fazla açıklayan etkileşimin korelasyon*model (kısmi η^2=0.706) olduğu ve diğer ikili etkileşimlerin etki büyüklerinin (kısmi η^2≤0.070) çok düşük olduğu gözlenmektedir. Üçlü ortak etkiler içerisinde yalnızca alt test sayısı*korelasyon*model ve alt test uzunluğu*korelasyon*model ortak etkileşimlerinin anlamlı fakat çok düşük düzeyde (kısmi η^2 ≤0.044) etkiye sahip olduğu görülmektedir. Dörtlü ortak etkinin güvenirlik değerlerinin varyansına anlamlı bir katkısı olmadığı görülmektedir (p=0.964). Çoklu karşılaştırma testi sonucunda ana etkiler için koşulların tüm düzeyleri arasında anlamlı farklılık olduğu bulunmuştur. Korelasyon, alt test uzunluğu ve alt test sayısı koşullarının düzeyleri arttıkça güvenirliğin arttığı gözlenmiştir. Modeller açısından en güvenilir kestirim yapan modeller sırasıyla Hiyerarşik ÇBMTK, Üst Düzey Sıralı ve İki Faktör modeldir.

Alt Test Puanlarına ait RMSE ve Güvenirlik Değerleri İçin Varyans Analizi Sonuçları

Simülasyon çalışmasında tüm koşullar altında elde edilen alt test puanlarına ait RMSE ve güvenirlik değerleri üzerinde model, alt test sayısı, alt test uzunluğu ve alt testler arasındaki korelasyonun etkisini incelemek için yapılan varyans analizi sonuçları sırasıyla Tablo 4 ve Tablo 5'te verilmiştir.

Tablo 4. Alt Test Puanlarına Ait RMSE Değerleri İçin Varyans Analizi Sonuçları

Varyans kaynağı	Kareler toplamı	df	Kareler ortalaması	F	p	Kısmi η^2
Alt test sayısı	0,002	1	,002	104,263	0,000	,029
Alt test uzunluğu	4,617	2	2,309	114331,070	0,000	,985
Korelasyon	5,680	3	1,893	93761,951	0,000	,988
Model	45,951	2	22,976	1137853,566	0,000	,998
Alt test sayısı * alt test uzunluğu	,000	2	,000	11,190	0,000	,006
Alt test sayısı * korelasyon	,101	3	,034	1662,740	0,000	,586
Alt test sayısı * model	,091	2	,045	2245,775	0,000	,560
Alt test uzunluğu * korelasyon	,092	6	,015	760,861	0,000	,564
Alt test uzunluğu * model	,425	4	,106	5257,361	0,000	,856
Korelasyon * model	16,664	6	2,777	137549,079	0,000	,996
Alt test sayısı * Alt test uzunluğu * korelasyon	0,001	6	,000	9,986	0,000	,017
Alt test sayısı * Alt test uzunluğu * model	0,000	4	,000	5,141	0,000	,006
Alt test sayısı * korelasyon * model	0,308	6	,051	2538,759	0,000	,812
Alt test uzunluğu * korelasyon * model	0,088	12	,007	364,595	0,000	,554
Alt test sayısı * Alt test uzunluğu * korelasyon * model	0,001	12	,000	4,206	0,000	,014

Tablo 4 incelendiğinde alt test puanlarına ait RMSE değerleri üzerinde tüm ana ve tüm ikili ortak etkilerin anlamlı düzeyde etkisi olduğu gözlenmektedir. Ana etkiler açısından etki büyüklükleri incelediğinde alt test puanlarına ait RMSE değerleri üzerinde alt test sayısı değişkeni (kısmi η^2 =0.029) dışındaki diğer değişkenlerin yüksek düzeyde (kısmi η^2≥0.985) etkiye sahip olduğu görülmektedir. İkili ortak etkiler incelendiğinde alt testlere ait RMSE değerlerinin varyansını en fazla açıklayan etkileşimin sırasıyla alt test uzunluğu* model (kısmi η^2=0.996) ve korelasyon*model (kısmi η^2=0.856) olduğu; en düşük etkiye sahip etkileşimin ise alt test sayısı*alt test uzunluğu (kısmi η^2=0.006) olduğu gözlenmektedir. Üçlü ortak etkiler içerisinde en fazla etkiye sahip etkileşimlerin sırasıyla alt test sayısı*korelasyon*model (kısmi η^2=0.812) ve alt test uzunluğu*korelasyon*model (kısmi η^2=0.554) olduğu görülürken diğer üçlü etkileşimlerin etkisinin çok düşük (kısmi η^2≤0.017) olduğu görülmektedir. Dörtlü etkilerin RMSE değerlerinin varyansına

katkısının çok düşük (kısmi η^2=0.014) olduğu görülmektedir. Çoklu karşılaştırma testi sonucunda ana etkiler için koşulların tüm düzeyleri arasında anlamlı farklılık olduğu bulunmuştur. Korelasyon arttıkça hatanın arttığı fakat alt test uzunluğu arttıkça hatanın azaldığı gözlenmiştir. İki ve üç boyutlu alt testlerden elde edilen alt test puanı kestirim hataları arasında bir fark olmadığı görülmüştür. Her bir boyut için elde edilen RMSE ortalaması sırasıyla 0,490 ve 0,491'dir. Varyans analinizde gözlenen anlamlı etkinin örneklem büyüklüğünden kaynaklandığı düşünülmektedir. Modeller açısından en az hatalı kestirim yapan modeller sırasıyla Üst Düzey Sıralı, Hiyerarşik ÇBMTK ve İki Faktör modeldir.

Tablo 5. Alt Test Puanlarına Ait Güvenirlik Değerleri İçin Varyans Analizi Sonuçları

Varyans kaynağı	Kareler toplamı	df	Kareler ortalaması	F	p	Kısmi η^2
Alt test sayısı	0,037	1	0,037	1487,220	0,000	0,297
Alt test uzunluğu	3,835	2	1,918	76934,827	0,000	0,978
Korelasyon	11,430	3	3,810	152859,783	0,000	0,992
Model	56,748	2	28,374	1138361,878	0,000	0,998
Alt test sayısı * alt test uzunluğu	0,000	2	0,000	2,827	0,059	0,002
Alt test sayısı * korelasyon	0,163	3	0,054	2175,031	0,000	0,649
Alt test sayısı * model	0,235	2	0,118	4724,061	0,000	0,728
Alt test uzunluğu * korelasyon	0,080	6	0,013	532,873	0,000	0,475
Alt test uzunluğu * model	0,115	4	0,029	1152,489	0,000	0,566
Korelasyon * model	28,721	6	4,787	192049,700	0,000	0,997
Alt test sayısı * Alt test uzunluğu * korelasyon	0,000	6	0,000	1,993	0,063	0,003
Alt test sayısı * Alt test uzunluğu * model	0,001	4	0,000	5,515	0,000	0,006
Alt test sayısı * korelasyon * model	0,450	6	0,075	3008,391	0,000	0,837
Alt test uzunluğu * korelasyon * model	0,059	12	0,005	197,838	0,000	0,402
Alt test sayısı * Alt test uzunluğu * korelasyon * model	0,001	12	0,000	2,238	0,008	0,008

Tablo 5 incelendiğinde alt test puanlarına ait güvenirlik değerleri üzerinde tüm ana etkilerin anlamlı düzeyde etkisi olduğu gözlenmektedir. Ana etkiler açısından etki büyüklükleri incelediğinde alt test puanlarına ait güvenirlik değerleri üzerinde alt test sayısı değişkeni (kısmi η^2=0.297) dışındaki diğer değişkenlerin yüksek düzeyde (kısmi η^2≥0.978) etkiye sahip olduğu görülmektedir. İkili ortak etkiler incelendiğinde alt testlere ait güvenirlik değerlerinin varyansına alt test sayısı*alt test uzunluğu etkileşiminin katkı sağlamadığı görülürken (p=0.059) en fazla katkı sağlayan etkileşimin korelasyon*model (kısmi η^2=0.997) olduğu ve diğer ikili etkileşimlerin etki büyüklerinin (kısmi η^2 ≥0.475) en az orta düzeyde olduğu görülmektedir. Üçlü ortak etkiler içerisinde yalnızca alt test sayısı*alt test uzunluğu*korelasyon etkileşimin anlamlı etkisinin olmadığı gözlenirken en fazla etkiye sahip üçlü ortak etkileşimin alt test sayısı*korelasyon*model (kısmi η^2=0.837) olduğu, en az etkiye sahip etkileşimin ise alt test sayısı*alt test uzunluğu*model (kısmi η^2=0.006) olduğu görülmektedir. Dörtlü etkinin alt testlere ait güvenirlik değerlerinin varyansına katkısının çok düşük düzeyde (kısmi η^2=0.997) olduğu görülmektedir. Çoklu karşılaştırma testi sonucunda ana etkiler için koşulların tüm düzeyleri arasında anlamlı farklılık olduğu bulunmuştur. Korelasyon arttıkça güvenirliğin azaldığı fakat alt test uzunluğu ve alt test sayısı arttıkça güvenirliğin arttığı gözlenmiştir. Modeller açısından en güvenilir kestirim yapan modeller sırasıyla Üst Düzey Sıralı, Hiyerarşik ÇBMTK ve İki Faktör modeldir.

Gerçek Veri Uygulamasına İlişkin Sonuçlar

Araştırmanın kapsamında "TEOG (2015) verilerinin Üst Düzey Sıralı (Higher Order), İki Faktör (Bi-factor) ve hiyerarşik çok boyutlu madde tepki kuramı modellerine göre alt test puan kestirimlerinin nasıl değiştiği" modellerin kestirdiği sonsal dağılımın ortalama ve standart sapması ve alt testler arası korelasyon değerleri ve standart sapması ile incelenmiştir. TEOG 2015 verisinin üç kestirim modeline göre her bir alt test için kestirilen sonsal dağılımın ortalama ve standart sapma değerleri Tablo 6'da verilirken kestirilen alt testler arası korelasyon matrisi ise Tablo 7'de verilmiştir.

Tablo 6. Sonsal Dağılımın Ortalama ve Standart Sapması

Alt testler	İki Faktör Model		Hiyerarşik ÇBMTK Model		Üst Düzey Sıralı Model	
	Ortalama	Std. Sapma	Ortalama	Std. Sapma	Ortalama	Std. Sapma
Din Kültürü	0.004	0.009	0.009	0.012	0.009	0.013
Fen Bilgisi	-0.008	0.014	-0.004	0.010	-0.001	0.009
İngilizce	-0.017	0.021	-0.026	0.028	-0.024	0.026
Matematik	-0.015	0.017	-0.017	0.019	-0.014	0.017
Tarih	-0.005	0.009	-0.005	0.009	-0.002	0.008
Türkçe	0.002	0.008	0.005	0.013	0.000	0.008

Tablo 7. Modellerin Kestirdiği ve TEOG Verisinin Alt Testler Arası Korelasyon Matrisi

Model	Alt testler	Din Kültürü	Fen Bilgisi	İngilizce	Matematik	Tarih	Türkçe
İki Faktör	Fen Bilgisi	-0.032	1	0.001	-0.009	-0.012	-0.007
	İngilizce	-0.037	0.001	1	-0.005	-0.018	-0.009
	Matematik	-0.036	-0.009	-0.005	1	0.000	-0.009
	Tarih	-0.037	-0.012	-0.018	0.000	1	0.006
	Türkçe	-0.041	-0.007	-0.009	-0.009	0.006	1
ÇBMTK	Fen Bilgisi	0.009	1	0.006	-0.007	-0.006	-0.005
	İngilizce	-0.008	0.006	1	-0.004	-0.012	-0.006
	Matematik	-0.005	-0.007	-0.004	1	0.005	-0.005
	Tarih	0.012	-0.006	-0.012	0.005	1	0.010
	Türkçe	-0.007	-0.005	-0.006	-0.005	0.010	1
Üst Düzey	Fen Bilgisi	0.008	1	0.006	-0.005	-0.005	-0.003
	İngilizce	-0.008	0.006	1	-0.004	-0.015	-0.010
	Matematik	-0.004	-0.005	-0.004	1	0.007	-0.007
	Tarih	0.015	-0.005	-0.015	0.007	1	0.013
	Türkçe	-0.006	-0.003	-0.010	-0.007	0.013	1
TEOG	Fen Bilgisi	-0.002	1	0.006	-0.005	-0.005	-0.003
	İngilizce	-0.018	0.013	1	-0.004	-0.015	-0.010
	Matematik	0.013	-0.011	-0.007	1	0.007	-0.007
	Tarih	0.017	0.011	-0.009	0.000	1	0.013
	Türkçe	0.040	-0.001	0.006	-0.081	0.010	1

Tablo 6 incelendiğinde İki Faktör modelin Hiyerarşik ÇBMTK modele göre Din Kültürü, İngilizce, Matematik ve Türkçe alt testlerinde Üst Düzey Sıralı modele göre ise Din Kültürü ve İngilizce alt testlerinde daha düşük sonsal standart sapmaya sahip olduğu görülmektedir. Üst Düzey Sıralı

modelin Hiyerarşik ÇBMTK modele göre Fen Bilgisi, İngilizce, Matematik, Tarih ve Türkçe alt testlerinde İki Faktör modele göre ise Fen Bilgisi ve Tarih alt testlerinde daha düşük sonsal standart sapmaya sahip olduğu görülmektedir. Hiyerarşik ÇBMTK modelin İki Faktör modele göre yalnızca Fen Bilgisi alt testinde ise Üst Düzey Sıralı modele göre ise yalnızca Din Kültürü alt testinde daha düşük sonsal standart sapmaya sahip olduğu görülmektedir. Tüm yöntemlerin bütün alt testler açısından sonsal dağılım ortalama ve standart sapmaları bir arada değerlendirildiğinde üç yönteminde genel olarak düşük sonsal standart sapmaya sahip olduğu fakat İki Faktör modelin diğer yöntemlere göre az farkla daha iyi sonuç verdiği söylenebilir.

Tablo 7'de üç model tarafından kestirilen alt testler arası korelasyon matrisi incelendiğinde üç modelin de gerçek veri matrisinin korelasyon matrisine çok benzer kestirimler yaptığı görülmektedir. Tablo 7'deki korelasyon matrisinin standart sapma değerleri ise EK-1'de verilmiştir. Standart sapma değerleriyle birlikte korelasyon değerleri incelendiğinde Hiyerarşik ÇBMTK model ile Üst Düzey Sıralı modelin birbirine çok benzer sonuçlar verdiği ve çok az farkla İki Faktör modelin diğer modellere göre standart sapma değerlerinin yüksek olduğu söylenebilir. Gerçek veri setinden elde edilen bu sonucun simülasyon çalışmasının 0.0 korelasyon ve 20 maddelik alt testlerden elde edilen sonuçları ile uyumludur.

SONUÇLAR ve TARTIŞMA

Araştırmada öncelikle iki ve üç boyutlu Hiyerarşik ÇBMTK'ya göre üretilen verilerin aynı modele dayalı elde edilen toplam ve alt test yetenek parametresi kestirimlerinin belirli bir hata düzeyinde elde edildiği gözlenmiştir. Simülasyon çalışmalarında verilerin belirli bir hata düzeyinde üretilmesi beklenen bir durumdur. RMSE istatistiği için belirli bir sınır olmadığından yalnızca daha düşük değerlerin daha iyi olduğu belirtilmektedir. Parametre doğrulama çalışmalarında yetenek parametresine ait hataların madde parametresine ait hatalardan daha yüksek elde edildiği görülmektedir (Çakıcı Eser, 2014; Jiang, Wang ve Weiss,2016; Lee, 2012). Ayrıca de la Torre ve Patz (2005) ile Yao'nun (2010) çok boyutlu MTK'ya dayalı ürettikleri verileri ve de la Torre, Song ve Hong'un (2011) Üst Düzey Sıralı Modele dayalı ürettikleri verileri aynı model ile analiz etmeleri sonucu yetenek parametre kestirimlerinde bu araştırma ile benzer düzeyde hataların elde edildiği görülmüştür. Bu durumun veri üretilirken yetenek parametresinin geniş bir normal dağılımdan gelmesi ve az sayıda madde örneklemi ile birey yeteneğinin kestirilmesinden kaynaklandığı düşünülmektedir.

Hem iki hem de üç boyutlu veri setlerinde alt testler arasındaki korelasyon düzeyinin artmasıyla Hiyerarşik ÇBMTK Model'den toplam test puanı için elde edilen hataların artmasının nedeni olarak testin boyutluluk derecesindeki azalma gösterilebilir. Bir başka ifadeyle alt testler arasındaki yüksek düzeyde ilişkiler testin tek boyutluluğa yakınlaşmasına neden olmaktadır. Aynı nedenle, İki Faktör Model'de toplam test puanı kestirimlerinde alt testler arasındaki korelasyon düzeyi arttıkça hataların azaldığı görülmektedir. Çünkü İki Faktör Model'de toplam test puanı testteki tüm maddelerin kalibrasyonundan elde edilir. Korelasyon düzeyindeki artışı ile Üst Düzey Sıralı Model'de toplam test puanı kestirimlerinde hataların azalmasının nedeni ise modelde kullanılan regresyon katsayılarının alt testler arası ilişkilerden türetilmesidir. Yukarıdaki sayılan benzer nedenler ile hem iki hem üç boyutlu verilerde İki Faktör ve Üst Düzey Sıralı Model için korelasyon arttıkça toplam test güvenirliği artmaktadır. Ayrıca alt testler arası korelasyonlar yüksek olsa da yapı modelinin maddeleri tek bir boyutla ilişkilendirmesi ve verilerin yine aynı modelle üretilmesi nedeniyle tüm koşullarda Hiyerarşik ÇBMTK Model en düşük hata ve en yüksek güvenilirlik düzeyinde sonuçlar vermektedir. Daha uzun alt testlerde üç yöntem için de kestirim hatalarının azalması ve güvenirliğin artması beklenen bir durumdur.

Hiyerarşik ÇBMTK Model'in toplam test puanı kestirimlerinde maksimum bilgi yöntemini kullanması nedeniyle alt test sayısındaki artışın yetenek parametre kestirimleri üzerinde en fazla katkı sağladığı model bu modeldir. Alt test sayısındaki artış toplam madde sayısını arttırdığı için İki Faktör Model'de toplam test yetenek kestirimlerine katkısı alt testler arası korelasyon arttıkça daha fazla artmaktadır. Üst Düzey sıralı modelde ise genel yetenek kestirimde kullanılan birinci düzey

değişken sayısının artması nedeniyle alt test sayısındaki artış kestirim hatalarını azaltmakta ve güvenirliği arttırmaktadır.

Hem iki hem de üç boyutlu veri setlerinde alt testler arasındaki korelasyon düzeyinin Hiyerarşik ÇBMTK Model'den elde edilen alt test puan kestirim hataları ve güvenirliği üzerinde bir etkisi olmadığı sonucuna ulaşılmıştır. Fakat Bulut (2013) ve Yao (2010) çalışmalarında alt testler arası korelasyon düzeyindeki artışın ÇBMTK Model kestirimlerinin güvenirliği ve doğruluğunu arttırdığını bulmuşlardır. Alt test sayısındaki artışın bu çalışmada kestirim hataları ve güvenirliği üzerinde bir etkisi gözlenmezken Bulut (2013) alt test sayısının güvenirlik üzerinde minimal düzeyde etkisi olduğunu belirtmiştir. Bu çalışma ile diğer çalışmalar arasındaki farkın veri üretme koşulları arasındaki farklılıktan kaynaklanabileceği düşünülmektedir.

Hem iki hem de üç boyutlu veri setlerinde alt testler arasındaki korelasyon düzeyindeki artış ile İki Faktör Model'den elde edilen alt test puan kestirim hatalarının arttığı ve güvenirliğin azaldığı sonucu Yao (2010) ve Chang'nın (2015) araştırma sonuçlarıyla uyumludur. Bu modelde alt testler arasındaki ilişkilerin dik olduğu varsayımı nedeniyle model yüksek ilişki gösteren alt testlerde yüksek hatalı ve düşük güvenirlikli kestirimler yapmaktadır. Dolayısıyla kabul edilebilir düzeyde güvenilir kestirimlerin ancak 0.0 korelasyon düzeyinde ve 0.3 korelasyon düzeyinin de uzun alt test düzeylerinde elde edilmektedir. Ayrıca alt test sayısındaki artışın bu model için optimal koşul olan düşük korelasyon düzeylerinde hataların azalması ve güvenirliğin artması benzer nedenlerden kaynaklanmakladır. Optimal koşullardan uzaklaştıkça alt test sayısını arttırmak daha düşük güvenirlikli ve daha yüksek hatalı kestirimlere sebep olmaktadır.

Hem iki hem de üç boyutlu veri setlerinde alt testler arasındaki korelasyon düzeyindeki artış ile Üst Düzey Sıralı Model'den elde edilen alt test puanı kestirim hataların azalması modelin doğası gereği genel ve alt boyutlar arasındaki ilişkileri kullanmasının doğal bir sonucudur. De la Torre, Song ve Hong (2011) Üst Düzey Sıralı Model'e dayalı ürettikleri veriler üzerinde de bu araştırmayla benzer sonuçlara ulaşmıştır. Aynı zamanda 0.8 korelasyon düzeyinde Üst Düzey Sıralı Model'in biraz daha doğru ve güvenilir kestirimler yapmasıyla birlikte bu model ile Hiyerarşik ÇBMTK model benzer performans göstermiştir. Bu durumun iki modelin de alt testler arasındaki korelasyonları kullanarak yetenek parametresi kestirmesinden kaynaklandığı düşünülmektedir. Bu araştırma sonuçlarından farklı olarak Yao (2010) bu iki modeli benzer ama minimal farkla ÇBMTK modelin daha iyi performans gösterdiğini bulurken de la Torre, Song ve Hong (2011) bu iki modelin performansını birbirine eşit bulmuştur. Bu araştırmanın alt test sayısındaki artış ile Üst Düzey Sıralı Model'in alt test puan kestirimlerinin hatası ve güvenirliği üzerinde minimal düzeyde iyileşme sağladığı sonucu ile de la Torre, Song ve Hong'un (2011) araştırma sonuçları benzedir. Daha uzun alt testlerde üç yöntem için de kestirim hatalarının azalması ve güvenirliğin artması beklenen bir durumdur.

Gerçek veri uygulamasında TEOG 2015 gerçek verisinin faktör analizi sonucunda elde edilen alt testler arası korelasyon matrisi ile Hiyerarşik ÇBMTK Model, Üst Düzey Sıralı Model ve İki Faktör Model tarafından kestirilen matris karşılaştırıldığında modellerin gerçek duruma çok benzer kestirimler yaptığı görülmüştür. Din Kültürü, Fen Bilgisi, İngilizce, Matematik, Tarih ve Türkçe alt testlerinin her bir yöntemden elde edilen sonsal dağılımın ortalama standart sapma değerleri karşılaştırıldığında üç yönteminde genel olarak düşük sonsal standart sapmaya sahip olduğu bulunmuştur. Fakat daha fazla alt düşük sonsal dağılım ortalama ve standart sapma değerleri verdiği için İki Faktör modelin diğer yöntemlere göre az farkla daha iyi sonuç verdiği sonucuna ulaşılmıştır.

Araştırmadan elde edilen bulgulara dayanarak; hiyerarşik modellerin varsayımının hem toplam test hem de alt test puanlarının kestiriminde farklı performans göstermesi nedeniyle daha doğru ve güvenilir alt test ve toplam test puanı kestirimleri için öncelikle mevcut testin yapısal modeli tespit edilmelidir. Toplam test puan kestirimlerinde araştırmada ele alınan tüm koşullar altında ve alt test puanların kestirimde ise hemen hemen tüm koşullarda en düşük hatalı ve en güvenilir kestirimlerin hiyerarşik ÇBMTK modelden elde edilmesi nedeniyle geniş ölçekli testlerin raporlanmasında bu modelin kullanımı önerilebilir. Alt testler arasında orta ve düşük düzeyde ilişkilerin olduğu bilinen sınavların raporlanmasında hiyerarşik ÇBMTK Model'e alternatif olarak bu modelle çok yakın analizler yapabilen Üst Düzey Sıralı Model'in kullanımı da tercih edilebilir. Alan yazında alt testler arasında yüksek düzeyde ilişki olduğu bilinen sınavlarda toplam test puanı kestirimleri için İki

Faktör Model'in kullanımı önerilirken bu araştırmada ele alınan koşullara sahip sınavlar için alt test puan kestirimlerinde bu yöntemin kullanımı önerilmez. Toplam test süresi göz ününde bulundurmak koşuluyla hem toplam hem de alt test puanları kestirimin doğruluğunu ve güvenirliğini arttıracağı için alt test uzunluklarının arttırılması önerilebilir.

KAYNAKÇA

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational, & Psychological Testing (US). (1999*). Standards for educational and psychological testing*. American Educational Research Association, Washington, DC.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika, 64*(2), 153–168, doi: 10.1002/j.2333-8504.1998.tb01752.x

Brandt, S., & Duckor, B. (2013). Increasing unidimensional measurement precision using a multidimensional item response model approach. *Psychological Test and Assessment Modeling, 55*(2), 148-161.

Brennan, R. L. (2012). *Utility indexes for decisions about subscores* (No. 33). Center for Advanced Studies in Measurement and Assessment (CASMA). Retrieved from https://education.uiowa.edu/sites/education.uiowa.edu/files/documents/centers/casma/publications/casma-research-report-33.pdf

Bulut, O. (2013). *Between-person and within-person subscore reliability: Comparison of unidimensional and multidimensional IRT models*. (Doctoral Dissertation). Retrieved from https://conservancy.umn.edu/bitstream/handle/11299/155592/Bulut_umn_0130E_13879.pdf?sequence=1&isAllowed=y

Chang, Y. F. (2015). *A Restricted Bi-factor Model of Subdomain Relative Strengths and Weaknesses.* (Doctoral Dissertation) Retrieved from https://conservancy.umn.edu/bitstream/handle/11299/175551/CHANG_umn_0130E_16452.pdf?sequence=1&isAllowed=y

Çakıcı Eser, D. (2015). *Çok boyutlu madde tepki kuramının farklı modellerinden çeşitli koşullar altında kestirilen parametrelerin incelenmesi.* (Doktora tezi). Erişim adresi: http://tez2.yok.gov.tr/

de la Torre, J., & Patz, R.J. (2005). Making the most of what we have: A practical application of multidimensional IRT in test scoring. *Journal of Educational and Behavioral Statistics, 30*(3), 295–311, doi: 10.3102/10769986030003295

de la Torre, J. (2009). Improving the quality of ability estimates through multidimensional scoring and incorporation of ancillary variables. *Applied Psychological Measurement, 33*(6), 465–485, doi: 10.1177/0146621608329890

de la Torre, J., & Song, H. (2009). Simultaneous estimation of overall and domain abilities: A higher-order IRT model approach. *Applied Psychological Measurement, 33*(8), 620-639, doi: 10.1177/0146621608326423

de la Torre, J., Song, H., & Hong, Y. (2011). A comparison of four methods of IRT subscoring. *Applied Psychological Measurement, 35(*4), 296-316, doi: 10.1177/0146621610378653

Edwards, M. C., & Vevea, J. L. (2006). An empirical Bayes approach to subscore augmentation: How much strength can we borrow?. *Journal of Educational and Behavioral Statistics, 31*(3), 241-259, doi: 10.3102/10769986031003241

ETS. (2014). *ETS standards for quality and fairness*. Educational Testing Service. .Retreived from https://www.ets.org/s/about/pdf/standards.pdf

Ferrara, S., & DeMauro, G. E. (2007). Standardized assessment of individual achievement in K–12. In R. L. Brennan (Eds.). *Educational measurement*, 579–622. Westport, CT: Praeger.

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2011). *How to design and evaluate research in education*. (8th edition). Boston: McGraw – Hill.

Gall M. D., Gall, J. P., & Borg, W., R. (2003). *Educational research: An introduction*. (7th. Edition). Pearson Education, Inc.

Gibbons, R. D., & Hedeker, D. (1992). Full-information item Bi-factor analysis. *Psychometrika, 57*, 423–436.

Haberman, S. J. (2008). When can subscores have value*? Journal of Educational and Behavioral Statistics, 33*(2), 204–229, doi:10.3102/1076998607302636

Haberman, S., Sinharay, S., & Puhan, G. (2009). Reporting subscores for institutions. *British Journal of Mathematical and Statistical Psychology, 62*(1), 79–95, doi:10.1348/000711007X248875

Haladyna, T. M., & Kramer, G. A. (2004). The validity of subscores for a credentialing Test. *Evaluation & The Health Professions, 27*(4), 349–368, doi: 10.1177/0163278704270010

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

Huang, H. Y., Wang, W. C., Chen, P. H., & Su, C. M. (2013). Higher-order item response models for hierarchical latent traits. *Applied Psychological Measurement, 37*(8), 619-637, doi: 10.1177/0146621613488819

Jiang, S., Wang, C., & Weiss, D. J. (2016). Sample size requirements for estimation of item parameters in the multidimensional graded response model. *Frontiers in psychology, 7*(109), 1-10, doi: 10.3389/fpsyg.2016.00109

Kelley, T. L. (1927). *The interpretation of educational measurements*. New York: World Book.

Kelley, T. L. (1947). *Fundamentals of statistics*. Cambridge: Harvard University Press Kerlinger.

Kerlinger, F.N. (1973). *Foundation of behavioural research*. New York. Holt. Rinehand and Hinston.

Köse, İ.A. (2010). *Madde tepki kuramına dayalı tek boyutlu ve çok boyutlu modellerin test uzunluğu ve örneklem büyüklüğü açısından karşılaştırılması*. (Doktora Tezi). Erişim adresi: http://tez2.yok.gov.tr/

Lee, J. (2012). *Multidimensional item response theory: an investigation of interaction effects between factors on item parameter recovery using Markov Chain Monte Carlo*. (Doctoral Dissertation). Retrieved from https://d.lib.msu.edu/islandora/object/etd:1577/datastream/OBJ/download/Multidimensional_item_response_theory__an_investigation_of_interaction_effects_between_factors_on_item_parameter_recovery_using_Markov_Chain_Monte_Carlo.pdf

Ling, G. (2012). *Why the major field test in business does not report subscores: Reliability and construct validity evidence* (No. RR-12-11). ETS Research Report. Retrieved from https://www.ets.org/Media/Research/pdf/RR-12-11.pdf

Lorenzo-Seva, U., & Ferrando, P.J. (2006). FACTOR: A computer program to fit the exploratory factor analysis model. *Behavioral Research Methods, Instruments and Computers, 38*(1), 88-91.

Messick, S. (1989). Validity. In R. L. Linn (Eds.). *Educational measurement*, 13-103, New York, NY: Macmillan.

Monaghan, W. (2006). The fact about subscores (No. RDC-04). ETS Research Report. Retrieved from https://www.ets.org/research/policy_research_reports/rdc-04

Özkan, Y. Ö. (2012). *Öğrenci başarılarının belirlenmesi sınavından (ÖBBS) klasik test kuramı, tek boyutlu ve çok boyutlu madde tepki kuramı modelleri ile kestirilen başarı puanlarının karşılaştırılması*. (Doktora Tezi). Erişim adresi: http://tez2.yok.gov.tr/

Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement*, 21, 25–36, doi: 10.1177/0146621697211002

Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika*, *22*(1), 53-61.

Sheng, Y., & Wikle, C. K. (2007). Comparing Multiunidimensional and unidimensional item response theory models. *Educational and Psychological Measurement, 67*(6) 899–919, doi: 10.1177/0013164406296977

Sheng, Y., & Wikle, C. K. (2008). Bayesian multidimensional IRT models with a hierarchical structure. *Educational and Psychological Measurement, 68*(3), 413–430, doi: 10.1177/0013164407308512

Shin, D. (2007). *A comparison of methods of estimating subscale scores for mixed-format tests*. Report for Pearson Educational Measurement. Retreived from https://images.pearsonassessments.com/images/tmrs/tmrs_rg/EstimatingSubscaleScoresforMixedFormatItemsforPEMreportfinal.pdf?WT.mc_id=TMRS_A_Comparison_of_Methods_of_Estimating

Shin, C. D., Ansley, T., Tsai, T., & Mao X. (2005, April). *A comparison of methods of estimating objective scores*. Annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Sinharay, S. (2010). How often do subscores have added value? Results from operational and simulated data. *Journal of Educational Measurement, 47*(2),150-174.

Skorupski, W. P., & Carvajal, J. (2010). A comparison of approaches for improving the reliability of objective level scores. *Educational and Psychological Measurement, 70*(3), 357-375, doi: 10.1177/0013164409355694

Wainer, H., Vevea, J. L., Camacho, F., Reeve, B. B., Rosa, K., Nelson, L.,Swygert, K. A., & Thissen, D. (2001). Augmented score—''borrowing strength'' to compute scores based on small numbers of items. In D. Thissen and H. Wainer (Eds.). *Test scoring*, (343-387). Mahwah, Lawrence Erlbaum Associates, Inc

Wang, W. C., & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement*, *29*(2), 126–149, doi: 10.1177/0146621604271053

Wang, W. C., Chen, P. H., & Cheng, Y. Y. (2004). Improving measurement precision of test batteries using multidimensional item response models. *Psychological Methods, 9*(1), 116, doi: 10.1037/1082-989X.9.1.116

Yao, L. (2003). SimuMIRT [Software]. Monterey, CA: Defense Manpower Data Center. Retreived from http://www.bmirt.com

Yao, L. (2010). Reporting valid and reliable overall scores and domain scores. *Journal of Educational Measurement, 47*(3), 339-360, doi: 10.1111/j.1745-3984.2010.00117.x

Yao, L. (2017). Comparing methods for estimating the abilities for the multidimensional models of mixed item types. *Communications in Statistics-Simulation and Computation*, 1-18, doi: 10.1080/03610918.2016.1277749

Yao, L., & Boughton, K. A. (2007). A multidimensional item response modeling approach for improving subscale proficiency estimation and classification. *Applied Psychological Measurement, 31*(2), 83-105, doi: 10.1177/0146621606291559

Yao, L., & Schwarz, R. D. (2006). A multidimensional partial credit model with associated item and test statistics: An application to mixed-format tests. *Applied Psychological Measurement, 30*(6), 469–492, doi: 10.1177/0146621605284537

Yen, W. M. (1980). The extent, causes and importance of context effects on item parameters for 2 latent trait models. *Journal of Educational Measurement, 17*(4), 297–311, doi: 10.1111/j.1745-3984.1980.tb00833.x

Yen, W. M. (1987,June). *A Bayesian/IRT index of objective performance*. Annual meeting of the Psychometric Society, Montreal, Quebec, Canada.

EXTENDED ABSTRACT

Introduction

In many developed countries, large-scale standardized tests are the most common measurement tools used in education and psychology. These tests have multiple components which consists of subsets of items that measure specific content or structure. Although the total test score estimate is useful for important decisions, the subscores complement the total test score estimate by providing finer grained diagnosis of weakness and strengths of examinees. In this context, there is an increasing interest in subscores in educational testing. Reliable subscores should be obtained to make valid inferences about attributes of a student in the subtests. In practice, conventional analysis of tests with multiple components ignores its multidimensionality and only responses specific to each subtest are used in estimating the subscores of examinees.

In this study, the relationship between subtest and total test was investigated by using hierarchical item response theory models in order to contribute to reliable subtest and total test score estimates. The RMSE and reliability of the total test score and subtest scores estimated by the Higher Order, Bi-factor and hierarchical MIRT models in the study were compared under the conditions of the size of the correlations between the subtest number, subtest length and number of subtests. In addition, the performance of three models used in the research was examined on TEOG 2015 data.

Method

To generate data sets based on the item parameters of the TEOG 2015 data, item discrimination parameters were drawn from normal distribution with a mean of 1.5 and a variance of 0.5; item difficulty parameters were drawn from normal distribution with a mean of 0.0 and a variance of 1.0, and guessing (lower asymptote) parameters were drawn from beta distribution with (6,16). The true subtest abilities were drawn from a multivariate normal distribution with variance-covariance matrix based on the correlations between the dimensions explained under simulation conditions. Finally, given subtest abilities and item parameters, binary responses were simulated for number of subtest (2,3), subtest length (20,30,40) and correlation between subtest (0.0, 0.3, 0.5, 0.8) by SimuMIRT software. The simulated data and TEOG 2015 data was analyzed by BMIRT software. For the parameter estimates, 3PL model and MCMC estimation method were used.

Results and Discussion

As a result of the study, when the correlation between the subtests and the subtest length increased, the RMSE of the ability parameters decreased and the reliability increased for the total test score obtained from Higher Order and Bi factor Models. But with higher levels of correlation between the subtests in both two- and three-dimensional datasets, more errors were obtained for total test score from the Hierarchical MIRT model. This might be caused by the decrease in the test dimensionality. For the same reason, it was observed that as the correlation level between subtests increased in the total test score estimates from the Bi Factor Model, the errors decreased since the total test score in the Bi Factor Model is obtained from the calibration of all the items in the test. As a result, it was found that Hierarchical MIRT Model outperformed Higher Order and Bi Factor Models with regards to total test score recovery for all conditions. But when the correlation between subtests was .8 all three methods performed similarly. Furthermore, the reliability values obtained from Hierarchical MIRT model under all conditions were at least .7. The increase in the number of subtests contributed to the more accurate total test score estimates for three models, with Hierarchical MIRT maximum information model performing slightly better than the two other methods.

Under all conditions, the lowest RMSE value and the highest reliability value were yielded from Hierarchical MIRT model for subscores recovery and Bi factor model performed the worst. When the correlation between the subtests increased in both two- and three-dimensional datasets, the RMSE of the ability parameters decreased and the reliability increased for the subscores obtained from Higher Order whereas RMSE of the ability parameters increased and the reliability decreased for the subscores obtained from Bi Factor model. These results for Bi Factor models are similar with the ones obtained from the paper of Yao (2010) and Chang (2015). Also, it was unexpectedly deduced that there was no effect of the level of correlation between the subtests on the subscores estimation errors and reliability obtained from the Hierarchical MIRT Model. However, it was found in the studies of Bulut (2013) and Yao (2010) that when the correlation between the subtests increased, the reliability and accuracy of the Hierarchical MIRT model subscores estimates increased. It was found from real data analysis that all three methods gave similar estimates for subscores. This was consistent with results obtained for the condition in which the correlation between subtests was .0.

According to the results of the research, to report total test scores with reliability greater than .8: The correlation between subtests has to be higher than .3 to use either Hierarchical MIRT or Higher Order model and has to be than .5 to use Bi factor model for a test of 20 items for each subtest. Also, the subtest length has to be at least 30 when the correlation between subtests is at .0 for Hierarchical MIRT model. Both Hierarchical MIRT and Higher Order model can give subscore estimates with greater than .8 at all level of correlation between subtests for at least a test of 20 items for each subtest.

Based on findings from the study; the use of the Hierarchical MIRT model is recommended for the reporting of large scale tests. In reporting exams known to have moderate and low correlations among the sub-tests, it may also be preferable to use the Higher Order model, which is able to perform close analyzes with the Hierarchical MIRT Model, as an alternative to the Hierarchical MIRT Model.

Ekler

Tablo EK-1. Modellerin Kestirdiği Alt Testler Arası Korelasyonların Standart Sapma Matrisi

Model	Alt testler	Din Kültürü	Fen Bilgisi	İngilizce	Matematik	Tarih	Türkçe
İki Faktör	Fen Bilgisi	0.033	1.000	0.007	0.011	0.014	0.009
	İngilizce	0.039	0.007	1	0.009	0.020	0.011
	Matematik	0.038	0.011	0.009	1	0.006	0.010
	Tarih	0.038	0.014	0.020	0.006	1	0.008
	Türkçe	0.041	0.009	0.011	0.010	0.008	1
ÇBMTK	Fen Bilgisi	0.010	1	0.009	0.009	0.008	0.007
	İngilizce	0.010	0.009	1	0.008	0.013	0.008
	Matematik	0.008	0.009	0.008	1	0.007	0.008
	Tarih	0.013	0.008	0.013	0.007	1	0.011
	Türkçe	0.009	0.007	0.008	0.008	0.011	1
Üst Düzey	Fen Bilgisi	0.010	1	0.010	0.009	0.008	0.007
	İngilizce	0.011	0.010	1	0.008	0.017	0.012
	Matematik	0.007	0.009	0.008	1	0.009	0.009
	Tarih	0.016	0.008	0.017	0.009	1	0.014
	Türkçe	0.008	0.007	0.012	0.009	0.014	1