Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi

Journal of Measurement and Evaluation in Education and Psychology

ISSN:1309-6575 *Yaz* 2020 *Summer* 2020 *Cilt: 11- Sayı: 2* *Volume: 11- Issue: 2*

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi

Journal of Measurement and Evaluation in Education and Psychology

ISSN: 1309 – 6575

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi (EPOD) yılda dört kez yayınlanan hakemli ulusal bir dergidir. Yayımlanan yazıların tüm sorumluğu ilgili yazarlara aittir.

İletişim
e-posta: epodderdergi@gmail.com
Web: https://dergipark.org.tr/tr/pub/epod

Journal of Measurement and Evaluation in Education and Psychology (EPOD) is a national refereed journal that is published four times a year. The responsibility lies with the authors of papers.

Contact
e-mail: epodderdergi@gmail.com
Web: http://dergipark.org.tr/tr/pub/epod

Owner

Dizinleme / Abstracting & Indexing
Emerging Sources Citation Index (ESCI), DOAJ (Directory of Open Access Journals), SCOPUS, TÜBİTAK TR DIZIN Sosyal ve Beşeri Bilimler Veri Tabanı (ULAKBİM), Tei (Türk Eğitim İndeksi)

Hakem Kurulu / Referee Board

Ahmet Salih ŞİMŞEK (Kırşehir Ahi Evran Üni.)
Ahmet TURHAN (American Institute Research)
Akif AVCU (Marmara Üni.)
Alperen YANDI (Abant İzzet Baysal Üni.)
Asiye ŞENGÜL AVŞAR (Recep Tayyip Erdoğan Üni.)
Ayfer SAYIN (Gazi Üni.)
Ayşegül ALTUN (Ondokuz Mayıs Üni.)
Arif ÖZER (Hacettepe Üni.)
Aylin ALBAYRAK SARI (Hacettepe Üni.)
Bahar ŞAHİN SARKIN (İstanbul Okan Üni.)
Belgin DEMİRUS (MEB)
Bengü BÖRKAN (Boğaziçi Üni.)
Betül ALATLI (Gaziosmanpaşa Üni.)
Betül TEKEREK (Kahramanmaraş Sütçü İmam Üni.)
Beyza AKSU DÜNYA (Bartın Üni.)
Bilge GÖK (Hacettepe Üni.)
Bilge BAŞUSTA UZUN (Mersin Üni.)
Burak AYDIN (Recep Tayyip Erdoğan Üni.)
Burcu ATAR (Hacettepe Üni.)
Burhanettin ÖZDEMİR (Siirt Üni.)
Celal Deha DOĞAN (Ankara Üni.)
Cem Oktay GÜZELLER (Akdeniz Üni.)
Cenk AKAY (Mersin Üni.)
Ceylan GÜNDEĞER (Aksaray Üni.)
Çiğdem REYHANLIOĞLU (MEB)
Cindy M. WALKER (Duquesne University)
Çiğdem AKIN ARIKAN (Ordu Üni.)
David KAPLAN (University of Wisconsin)
Deniz GÜLLEROĞLU (Ankara Üni.)
Derya ÇAKICI ESER (Kırıkkale Üni)
Derya ÇOBANOĞLU AKTAN (Hacettepe Üni.)
Devrim ALICI (Mersin Üni.)
Devrim ERDEM (Niğde Ömer Halisdemir Üni.)
Didem KEPIR SAVOLY
Didem ÖZDOĞAN (İstanbul Kültür Üni.)
Dilara BAKAN KALAYCIOĞLU (Gazi Üni.)
Dilek GENÇTANRIM (Kırşehir Ahi Evran Üni.)
Durmuş ÖZBAŞI (Çanakkele Onsekiz Mart Üni)
Duygu Gizem ERTOPRAK (Amasya Üni)
Duygu KOÇAK (Alanya Alaaddin Keykubat Üni.)
Ebru DOĞRUÖZ (Çankırı Karatekin Üni.)
Elif Bengi ÜNSAL ÖZBERK (Trakya Üni.)
Emine ÖNEN (Gazi Üni.)
Emrah GÜL (Hakkari Üni.)
Emre ÇETİN (Doğu Akdeniz Üni.)
Emre TOPRAK (Erciyes Üni.)
Eren Can AYBEK (Pamukkale Üni.)
Eren Halil ÖZBERK (Trakya Üni.)
Ergül DEMİR (Ankara Üni.)
Erkan ATALMIS (Kahramanmaras Sütçü İmam Üni.)
Ersoy KARABAY (Kirşehir Ahi Evran Üni.)
Esin TEZBAŞARAN (İstanbul Üni)
Esin YILMAZ KOĞAR (Niğde Ömer Halisdemir Üni.)
Esra Eminoğlu ÖZMERCAN (MEB)
Fatih KEZER (Kocaeli Üni.)
Fatih ORCAN (Karadeniz Teknik Üni.)
Fatma BAYRAK (Hacettepe Üni.)
Fazilet TAŞDEMİR (Recep Tayyip Erdoğan Üni.)
Fulya BARIŞ PEKMEZCİ (Bozok Üni.)
Funda NALBANTOĞLU YILMAZ (Nevşehir Üni.)
Gizem UYUMAZ (Giresun Üni.)
Gonca USTA (Cumhuriyet Üni.)
Gökhan AKSU (Adnan Menderes Üni.)
Gözde SIRGANCI (Bozok Üni.)
Gül GÜLER (İstanbul Aydın Üni.)
Gülden KAYA UYANIK (Sakarya Üni.)
Gülşen TAŞDELEN TEKER (Hacettepe Üni.)
Hakan KOĞAR (Akdeniz Üni.)
Hakan SARIÇAM (Dumlupınar Üni.)
Hakan Yavuz ATAR (Gazi Üni.)
Halil İbrahim SARI (Kilis Üni.)
Halil YURDUGÜL (Hacettepe Üni.)
Hatice KUMANDAŞ (Artvin Çoruh Üni.)
Hülya KELECİOĞLU (Hacettepe Üni.)
Hülya YÜREKLI (Yıldız Teknik Üni.)
İbrahim Alper KÖSE (Abant İzzet Baysal Üni.)
İlhan KOYUNCU (Adıyaman Üni.)
İlkay AŞKIN TEKKOL (Kastamonu Üni.)
İlker KALENDER (Bilkent Üni.)
İsmail KARAKAYA (Gazi Üni.)
Kübra ATALAY KABASAKAL (Hacettepe Üni.)
Levent YAKAR (Kahramanmaraş Sütçü İmam Üni.)
Mehmet KAPLAN (MEB)
Melek Gülşah ŞAHİN (Gazi Üni.)
Meltem ACAR GÜVENDİR (Trakya Üni.)
Meltem YURTÇU (İnönü Üni.)
Metin BULUŞ (Adıyaman Üni)
Murat Doğan ŞAHİN (Anadolu Üni.)
Mustafa ASİL (University of Otago)
Mustafa İLHAN (Dicle Üni.)
Nagihan BOZTUNÇ ÖZTÜRK (Hacettepe Üni.)
Nail YILDIRIM (Kahramanmaraş Sütçü İmam Üni.)
Neşe GÜLER (İzmir Demokrasi Üni.)
Neşe ÖZTÜRK GÜBEŞ (Mehmet Akif Ersoy Üni.)
Nuri DOĞAN (Hacettepe Üni.)
Nükhet DEMİRTAŞLI (Emekli Öğretim Üyesi)
Okan BULUT (University of Alberta)
Onur ÖZMEN (TED Üniversitesi)
Ömer KUTLU (Ankara Üni.)
Ömür Kaya KALKAN (Pamukkale Üni.)
Önder SÜNBÜL (Mersin Üni.)
Özge ALTINTAS (Ankara Üni.)

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi (Haziran 2020, Sayı: 11-2)

Journal of Measurement and Evaluation in Education and Psychology (June 2020, Issue: 11-2)

İÇİNDEKİLER / CONTENTS

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2020; 11(2); 98-112

Parametrik ve Parametrik Olmayan Madde Tepki Kuramında Farklı Örneklem Büyüklüklerine ve Boyutluluklarına Göre Parametre Değişmezliğinin İncelenmesi *

Çiğdem REYHANLIOĞLU ** Nuri DOĞAN ***

Öz
Bu çalışmanın amacı farklı boyutluluk ve örneklem özelliklerinde Madde Tepki Kuramı (MTK) uygulamalarına göre kestirilen parametrelerin değişmezliğini incelemektir. Bu amaçla 2015 yılındaki Temel Eğitimden Ortaöğretime Geçiş (TEOG) Sistemi'nin birinci uygulamasındaki A kitapçığını alan öğrenci cevapları araştırma verisi olarak kullanılmıştır. Çalışma evreninin büyüklüğü 63,871'dir. Evrenden rastgele seçilmiş 50, 100, 200, 500, 1000 ve 5000 kişilik gruplar çalışmanın örneklemini oluşturmaktadır. MTK uygulamalarında tek boyutlu Matematik alt testinden ve yapay olarak oluşturulan iki boyutlu testin sonuçlarından yararlanılmıştır. Çalışmadan elde edilen bulgular sonucunda tek boyutlu test için Tek Boyutlu Parametrik Olmayan MTK'ye (TBPOMTK) göre 200 örneklem büyüklüğü itibariyle madde parametresi değişmezliği sağlanmıştır. Aynı test Tek Boyutlu Parametrik MTK'ye (TBPMTK) göre analiz edildiğinde ise evren değere yakın madde parametresi kestirimleri için en az 1000 örneklem büyüklüğü ile çalışılması gerektiği sonucuna ulaşılmıştır. İki boyutlu testin TBPMTK ve TBPOMTK'ye ve Çok Boyutlu MTK'ye (ÇBMTK) göre analiz edilmesi ile elde edilen madde parametrelerinde değişmezliğin sağlanamadığı sonucuna ulaşılmıştır.

Anahtar Kelimeler: Madde Tepki Kuramı uygulamaları, örneklem büyüklüğü, parametre değişmezliği.

GİRİŞ

Alanyazında bir testin çok boyutlu olduğu bir durumda bireylerin performansını belirlemede tüm teste ait puanının kullanılıp kullanılamayacağı veya nasıl kullanılacağı tartışılmaktadır. Dolayısıyla bir testin tek boyutlu olmadığı bir durumda tekboyutluluk varsayımı üzerine kurulan test teorileri çok boyutlu testlerden elde edilen verilerin analizinde yetersiz kalabilir. Bu durumda çok boyutlu bir testten elde edilen bireylere ait yetenek ve madde parametrelerinin kestirilmesi sürecinde kullanılan modeller önemlidir (Meara, Robin ve Sireci, 2000). Bu modeller tek boyutluluk varsayımı gerektirmeyen çok boyutlu verileri analiz edebilecek nitelikteki modeller olmalıdır. Diğer yandan ölçme kuramları incelendiği zaman bir çok varsayım gerektiren parametrik yöntemlerin kullanılması oldukça yaygındır. Ancak parametrik koşulların oluşamadığı ve tek boyutluluğun sağlanmadığı durumlara eğitim uygulamalarında sık sık rastlanmaktadır. Eğitimde ve psikolojide her zaman parametrik koşulların oluşmaması ve bireyler hakkında karar vermek için birden fazla alt boyutu olan tstlerin kullanma zorunluluğu parametrik ve tek boyutlu modellerden farklı uygulamaların geliştirilmesini zorunlu kılmıştır. Bir başka ifade ile parametrik koşulların karşılanmadığı durumlarda parametrik olmayan modeller ve tek boyutluluğun sağlanamadığı durumlar için ise çok boyutlu modeller geliştirilmiştir. Ancak geliştirilen bu modeller kullanılmadan önce işlerliğinin deneysel olarak ortaya konması gerekir. Bu amaçla kuramlar çerçevesinde geliştirilen bu modellerden elde edilen sonuçların, mevcut kuramlardan elde edilen sonuçlarla karşılaştırılması gerekir.

* Bu çalışma, Parametrik ve Parametrik Olmayan Madde Tepki Kuraminda Farkli Örneklem Büyüklüklerine ve Boyutluluklarina Göre Parametre Değişmezliğinin İncelenmesi isimli doktora tezinden üretilmiştir.

** Dr., Gaziantep Kolej Vakfı Özel Okulları, Gaziantep-Türkiye, dr.cigdemreyhanlioglu@gmail.com: ve ORCID ID: 0000-0002-4685-0495
*** Prof.Dr., Hacettepe Üniversitesi, Ankara-Türkiye, e-posta: nuridogan2004@gmail.com ve ORCID ID: 0000-0001-6274-2016

Bu makaleye atıfta bulunmak için:
Reyhanlıoğlu, Ç., & Doğan, N. (2020). Parametrik ve parametrik olmayan madde tepki kuramında farklı örneklem büyüklüklerine ve boyutluluklarına göre parametre değişmezliğinin incelenmesi. Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, *11*(2), 98-112. doi: 10.21031/epod.584977

Geliş Tarihi:01.07.2019
Kabul Tarihi: 07.04.2020

Parametrik MTK modellerinin geliştirilmesi ve yaygınlaşması modern test teorisinin geliştirilmesinde ve kullanılmasında kuşkusuz önemli bir aşamadır. Bu modellerin kullanılması için büyük örneklemlere ihtiyaç duyulduğuna dair önemli bulgular elde edilmiştir. Bu durum parametrik modellerin okullarda ve diğer saha çalışmalarında uygulanabilirliği açısından önemli bir sınırlıktır. Bu sınırlıktan dolayı saha çalışmalarında ve özellikle okullarda madde parametreleri örneklem özelliklerine bağlı olarak, bireylerin başarı düzeyleri de madde parametrelerine bağlı olarak kestirilebilmektedir. Oysa MTK'nin en önemli özelliği olan madde ve yetenek parametrelerinin birbirinden bağımsız olarak kestirilmesini ifade eden değişmezlik özelliğinin sağlanmasının okullarda öğrenci yetenek düzeyleri ile test ve madde parametrelerine dayalı olarak alınan kararlar açısından önemli olduğu düşünülmektedir. MTK'nin değişmezlik özelliğinin sağlanması ile okullarda uygulanan testlere ait madde parametrelerinin öğrencilerin yetenek dağılımından bağımsız ve benzer şekilde yetenek dağılımlarının da test ve madde parametrelerinden bağımsız kestirilmesi mümkün olur (Price, 2017). Böylece değişmezlik özelliği sayesinde okullarda uygulanan sınavlara bağlı olarak alınan kararların daha isabetli olması beklenir. Bu nedenle küçük çalışma gruplarına uygulanabilmesi açısından özellikle okullarda parametrik olmayan MTK modellerinin kullanılması ve yaygınlaştırılması büyük önem arzetmektedir. Büyük örneklemlerin kullanılması zorunluluğuna bir çözüm sunması açısından parametrik olmayan MTK modelleri parametrik modellere göre önemli avantajlara sahiptir. Parametrik olmayan MTK modellerinin parametrik modellere göre bir diğer önemli avantajı da maddelere verilen tepkiler ile maddelerin ölçtüğü gizil değişken arasındaki ilişkinin daha az varsayıma sahip olmasıdır. Bunun nedeni parametrik olmayan MTK modellerine ait madde karakteristik eğrilerinin önceden tanımlanmış parametrik bir biçimlerinin olmamasıdır (Sodano & Tracey, 2011). Buna göre parametrik olmayan modellerin parametrik modellere göre daha kullanışlı olduğu ifade edilebilir. Bununla birlikte her nekadar parametrik olmayan modeller önemli avantajlara sahip olsa da bu modelleri kullanabilmek için en az parametrik modeller kadar iyi çalıştığına ilişkin kanıt toplamaya ihtiyaç vardır ve bu çalışmadan elde edilen sonuçlar bu açıdan önemli kanıtlar ortaya koyacaktır. Bu çalışmada Türkiye'de kullanımı yeni yaygınlaşan bir kuram tek boyutlu parametrik olmayan MTK'den elde edilen sonuçlar ile bu kuramın parametrik karşılığı tek boyutlu MTK ve çok boyutlu testlerin analizide kullanılan çok boyutlu MTK kapsamında elde edilen sonuçlar karşılaştırılmıştır.

Araştırmanın Amacı

Türkiye'deki alanyazın incelendiğinde, çok kategorili veriler üzerinde parametrik olmayan tek boyutlu ve çok boyutlu MTK uygulamalarının karşılaştırılması ile ilgili farklı çalışmalar olmasına rağmen (Koğar, 2018; Şengül-Avşar, 2018; Şengül-Avşar, 2017) iki kategorili veriler ile gerçekleştirilen sınırlı sayıda çalışmaya rastlanmıştır. Koğar (2014) ve Mor-Dirlik (2017) tarafından yapılan çalışmalar dışında araştırmaya ulaşılamamıştır. Bu çalışma, iki kategorili gerçek veri üzerinden gerçekleştirilen Türkiye literatüründeki ilk çalışmalardan biri olması bakımından önemlidir. Bu bağlamda bu çalışmada iki kategorili tek boyutlu bir veri seti üzerinden parametrik ve parametrik olmayan MTK modellerinin işlerliğinin test edilmesi amaçlanmıştır. Bununla birlikte çok boyutlu bir veri setinin tek boyutlu parametrik ve parametrik olmayan MTK modelleri ile çok boyutlu MTK modelinden elde edilen sonuçları karşılaştırılmıştır. Böylece tek boyutluluk varsayımının bozulduğu bir durumda tek boyutlu parametrik ve parametrik olmayan modellerin verinin analizinde yeterliliğinin ortaya konması amaçlanmıştır.

Güvenilir bilginin tekrarlanabilir bir yapıya sahip olması beklenir. Dolayısıyla aynı nedenlerin aynı koşullar altında aynı sonuçları vermesi beklenir. Yinelenmeyen, neden-sonuç ilişkisine dayanmayan bilgiler bilimsel sayılmayabilir. Bilimsel bilgi, incelenen alanların değişik ortamlarda tekrarlanmasıyla elde edilen sonuçlarından oluşur. Aynı zamanda bilimsel bilgi birikimlidir. Her bilim insanının yaptığı bir çalışmanın, kendinden önceki çalışmaları destekler nitelikte olması beklenir; desteklemediği durumda ise nedenlerinin açıkça ortaya konması beklenir. Bu bağlamda Türkiye ve dünya literatürü incelendiği zaman MTK uygulamalarından elde edilen sonuçların, ölçeklerdeki boyut sayısı (Smits,

Timmerman & Meijer 2012), yetenek düzeylerinin dağılımı (Syu, 2013) gibi farklı değişkenler açısından karşılaştırıldığı çok sayıda çalışmanın olduğu görülmüştür. Bu değişkenlerden bir tanesi de örneklem büyüklüğüdür (Koğar, 2014; Köse, 2010; Sünbül, 2011). Örneklem büyüklüğü, aynı amaca hizmet eden parametrik ve parametrik olmayan modellerden hangisinin kullanılması gerektiğine karar verirken göz önünde bulundurulması gereken önemli bir faktördür. Bu çalışmada elde edilen sonuçların farklılaşıp farklılaşmadığı kuramsal uygulamaların yanı sıra örneklem büyüklüğüne göre de incelenmiştir. Bu bağlamda bu çalışma, MTK uygulamalarından elde edilen sonuçların örneklem büyüklüğüne göre karşılaştırıldığı çalışmalardan biri olması açısından da önemlidir. Bu çalışma ile alanyazında parametrik ve parametrik olmayan MTK modellerinin kullanım tercihi için ölçüt olabilecek bir örneklem büyülüğü belirlemek çalışmanın amaçlarından biridir. Bu amaç doğrultusunda TEOG uygulamasının farklı boyutlardaki alt testlerinden elde edilen yetenek düzeylerinin, örneklem büyüklükleri de göz önünde bulundurularak parametrik ve parametrik olmayan MTK ile çok boyutlu MTK kapsamında farklılaşıp farklılaşmadığını ortaya koymak amaçlanmıştır. Bu amaç çerçevesinde cevap aranan problem cümlesi "Farklı boyutluluk ve örneklem büyüklüğü değişkenlerine göre parametrik ve parametrik olmayan MTK için parametrelerin değişmezliği ne düzeyde sağlanmaktadır?" şeklinde yapılandırılmıştır. Yapılandırılan problem cümlesi çerçevesinde cevap aranan alt problemler:

1. Tek boyutlu testlerde, parametrik ve parametrik olmayan MTK'ye göre hesaplanan madde parametreleri için kestirilen standart hata ortalamaları, örneklem büyüklüğü 50, 100, 200, 500, 1000 ve 5000 olduğunda ve evrenden kestirildiği durumda nasıldır?
2. Çok boyutlu testlerde, tek ve çok boyutlu MTK'ye göre hesaplanan madde parametreleri için kestirilen standart hata ortalamaları, örneklem büyüklüğü 50, 100, 200, 500, 1000 ve 5000 olduğunda ve evrenden kestirildiği durumda nasıldır?
3. Çok boyutlu testlerde, parametrik ve parametrik olmayan MTK'ye göre hesaplanan madde parametreleri için kestirilen standart hata ortalamaları, örneklem büyüklüğü 50, 100, 200, 500, 1000 ve 5000 olduğunda ve evrenden kestirildiği durumda nasıldır?

Alt problemlerin çözümünden elde edilen bulguları yorumlarken TBPOMTK, TBPMTK ve ÇBMTK için madde ayırtediciliği ve madde güçlüğü için farklı göstergelerden yararlanılır. TBPOMTK için maddenin ayırtedicilik gücünü yorumlarken Hi parametresi kullanılmıştır ve Hi değerinin 0.30'dan küçük olması Sijtsma ve Molenaar'a göre (2002) maddenin ayırtedicilik bakımından zayıf olduğunu gösterir. TBPOMTK için madde güçlüğünün göstergesi olarak klasik güçlük parametresi olan p değerlerinden yararlanılmıştır. TBPMTK için ise ayırt ediciliğin göstergesi olarak a parametresi, madde güçlüğünün göstergesi olarak b parametresi kullanılır. Teorik olarak a ve b parametreleri ($-\infty$, $+\infty$) aralığında değerler alırlar. Son olarak ÇBMTK için, TBPMTK'de olduğu gibi ayırt ediciliğin göstergesi olarak a parametresi kullanılır. ÇBMTK'de testin her bir boyutu için ayrı bir ayırtedicilik parametresi kullanılmaktadır. Bu çalışmada kullanılan birleşik test iki boyutlu olduğu için iki tane ayırtedicilik parametresi kestirilmiştir. Bunlar a_1 ve a_2 parametreleridir. ÇBMTK için madde güçlüğünün göstergesi olarak d parametresi kullanılır. Benzer şekilde d parametresi TBPMTK'deki b parametresi gibi yorumlanmaktadır.

YÖNTEM

Bu çalışma tek boyutlu parametrik ve parametrik olmayan MTK ile çok boyutlu MTK modellerine ilişkin betimleyici istatistikler elde etme, iki ya da daha fazla değişken arasındaki ilişkinin varlığını ve derecesini ortaya koyma açısından betimsel bir çalışmadır. Betimsel araştırmalar, olayların, objelerin, varlıkların, kurumların ve çeşitli alanların "ne" olduğunu açıklamaya çalışır (Kaptan 1977).

Evren ve Örneklem

Çalışmada kullanılan veriler, 2015 yılındaki TEOG'un birinci uygulamasında yer alan testlerin her biri için A kitapçığını alan öğrenci cevaplarından elde edilmiştir. A kitapçığında Türkçe, Matematik, Fen

Bilgisi, Din Kültürü ve Ahlak Bilgisi, İnkılap Tarihi, İngilizce, Almanca ve Fransızca alt testleri yer almaktadır. Bu çalışmada bütün alt testlerde A kitapçığında yer alan soruları cevaplayan öğrenciler çalışmanın evrenini oluşturmaktadır. Çalışmanın bundan sonraki kısmında calışma evreni "evren" şeklinde ifade edilmiştir. Evren büyüklüğü 63.871'dir. Bu çalışmada elde edilen sonuçlar örneklem büyüklüklerine göre karşılaştırıldığı için, evrenden rastgele seçilmiş 50, 100, 200, 500, 1000 ve 5000 kişilik gruplar çalışmanın örneklemlerini oluşturmaktadır. Örneklem büyüklükleri belirlenirken özellikle küçük örneklemler için saha uygulamaları göz önünde bulundurulmuştur. Örneklem büyüklüğünün alt sınırı belirlenirken 2017-2018 eğitim ve öğretim yılı içerisinde Gaziantep'teki özel kurumlarda okumakta olan ortalama 8. sınıf öğrenci sayısı göz önünde bulundurulmuştur. Gaziantep İl Milli Eğitim Müdürlüğü'ne bağlı olan Ar-Ge biriminden alınan bilgiye göre Gaziantep'te bulunan 24 özel ortaokulda okumakta olan toplam 1188 tane sekizinci sınıf öğrencisi bulunmaktadır. Dolayısıyla her bir okul başına düşen öğrenci sayısının ortalama değeri 49.5'tir. Bu nedenle örneklem büyüklüğünün alt sınırı 50 olarak belirlenmiştir. Diğer küçük örneklem büyüklükleri (100, 200 ve 500) de 50'nin çift katları olacak şekilde belirlenmiştir. Büyük örneklem büyüklerinin belirlenmesinde ise Hullin, Lissak ve Drasgow (1982), Goldman ve Raju (1986) ve Thissen ve Wainer (1982) tarafından yürütülen çalışma sonuçları göz önünde bulundurulmuştur.

Evrenden rastgele seçilen 50, 100, 200, 500, 1000 ve 5000 kişilik örneklemler replikasyon yapılmaksızın sadece bir kez seçilmiştir. Örneklem seçiminde replikasyon yapılmaması çalışmanın sınırlılığı gibi görülmesine rağmen büyük örneklemler için replikasyon yapılması durumunda karşılaşılacak problemlerin üstesinden gelmek için bu yola başvurulmuştur. Örnek olarak 5000 kişilik bir örneklem seçerken 50-100 replikasyon yapıldığında her bir örneklemde çok sayıda aynı birey bulunabilir ve bu bireyler parametre değişmezliğini şişirebilir. Bir başka ifadeyle değişmezliğin sağlanmasına otomatik neden olurlar ve/veya kestirimlerin yanlı olmasına neden olabilirler. Bu nedenle bu çalışmada replikasyon yapılmamıştır.

Veri Toplama Araçları

Bu çalışmada 2015 yılı TEOG birinci sınavının A kitapçığında yer alan Türkçe, Matematik, Fen ve Teknoloji, T.C. İnkılap Tarihi, Yabancı Dil, Din Kültürü ve Ahlak Bilgisi alt testlerinden elde edilen verilerden yararlanılmıştır.

İşlem

Çalışmanın amacına uygun olarak TEOG'un birinci uygulamasında kullanılan A kitapçığında yer alan tek boyutlu ve iki boyutlu olan alt testlere ait sonuçların analiz sürecinde kullanılması amaçlanmıştır. Bu amaçla TEOG'da yer alan bütün alt testlerin KMO ve Bartlett Küresellik Testi sonuçlarına göre faktör analizine uygun olup olmadığı incelenmiştir. TEOG'da yer alan bütün alt testlere ait KMO değerleri 0.90'ın üzerinde çıkmıştır. Bartlett Küresellik Testi sonuçları ise bütün alt testler için istatistiksel olarak anlamlıdır. Bu durumda TEOG'da yer alan her bir alt test için elde edilen Bartlett testinin istatiksel olarak anlamlı olması, verilerin çok değişkenli normal dağılımdan geldiğini ve dolayısıyla verilerin, faktör analizinin uygulanması için uygun bir yapıya sahip olduğunu gösterir (Çokluk, Şekercioğlu & Büyüköztürk, 2010). Bunun yanında örneklem büyüklüğünün faktör analizine uygunluğu açısından KMO değerinin 0.60'dan büyük olması istenir (Tabachnick & Fidell, 2001). Elde edilen sonuçlar doğrultusunda her bir alt test için polichoric korelasyon matrisinin kullanıldığı paralel analize dayalı boyutluluk sonuçları veren FACTOR 10.5.01 programından yararlanılarak faktör analizi gerçekleştirilmiştir. Yapılan boyutluluk analizi sonucunda TEOG'un bütün alt testlerinin baskın bir tek boyuta sahip olduğu belirlenmiştir. Her bir alt teste ait verinin, elde edilen tek faktörlü modelle olan uyumunu ortaya koymak için iki göstergeden yararlanılmıştır. Bunlar GFI ve RMSR'dir. GFI'nın 1'e yakınlığı ölçüsünde model ile veri uyumludur. RMSR ise Kelly'nin ölçüt değeri olan (0.0316)'dan küçük olursa model ile verinin iyi uyum sağladığı ifade edilebilir (Harman, 1962). Elde edilen GFI değerleri [0.998, 1] aralığında ve RMSR değerleri ise [0.001, 0.023] aralığında değerler almıştır.

TEOG'da yer alan alt testlerin tamamının tek boyutlu bir yapıya sahip olduğu belirlenmesinin ardından ÇBMTK analizleri için kullanılmak üzere, iki tane alt testten madde seçerek iki boyutlu ve tek boyutlu MTK analizlerinde kullanılan testle aynı uzunluğa sahip (20 madde) yeni bir test oluşturulmuştur. ÇBMTK analizleri için gerekli olan testin oluşturulmasında yararlanılan alt testlerin seçiminde alt testlerin ikili kombinasyonlarından elde edilen faktör analizi sonuçları göz önünde bulundurulmuştur. Birleştirildiğinde iki boyutluluğu en iyi sağlayan alt testler Fen Bilgisi ile Din Kültürü ve Ahlak Bilgisi alt testleridir. Fen Bilgisi ile Din Kültürü ve Ahlak Bilgisi alt testlerinden kendi içlerinde korelasyonları yüksek ve diğer test maddeleriyle korelasyonları düşük 10'ar maddenin seçilmesiyle 2 boyutlu birleşik bir test oluşturulmuştur. Oluşturulan testin 2 boyutlu yapısına dair gerekli kanıtlar faktör analizi ile elde edilmiştir. Mevcut verinin, elde edilen iki faktörlü modelle olan uyumunu ortaya koymak için kullanılan iki gösterge arasından GFI değeri 0.999 ve RMSR değeri ise 0.01'dir. Sonuçta birleşik teste ait verilerin 2 faktörlü yapıya uyum sağladığı sonucuna ulaşılmıştır.

Verilerin Analizi

Verilerin analizi süreci MTK'nin gerekli varsayımlarının test edilmesiyle başlamıştır. TBPMTK için tek boyutluluk ÇBMTK için yerel bağımsızlık ve TBPOMTK için tek boyutluluk ve monotonluk varsayımları test edilmiştir. TBPMTK ve TBPOMTK için yerel bağımsızlık varsayımı ayrıca test edilmemiştir. Bunun nedeni tek boyutluluğun sağlanmasının yerel bağımsızlık varsayımının sağlanması için yeterli görülmektedir (Hambleton, Swaminathan ve Rogers, 1985). Bu nedenle yerel bağımsızlık varsayımı yalnızca ÇBMTK için test edilmiştir.

Yerel bağımsızlığı test etmek için kullanılan yöntemlerden bir tanesi koşullu maddeler arası korelasyonların incelenmesidir (Ferrara, Huyny&Baghi, 1997; Akt: Bulut, 2015). Bu çalışmada yerel bağımsızlığı test etmek için belli bir yetenek ranjı arasındaki (yüksek yetenek ve düşük yetenek grupları) maddelerarası korelasyonlardan yararlanılmıştır. Alt grup ile üst grupları belirlemek için ham puanların %20 ve %80'lik dilimleri kullanılır. Sınırlı yetenek düzeyindeki bireyler için elde edilen varyans ve kovaryans veya korelasyon matrislerinin köşegeninde yer alan elementlerin 0 veya 0'a çok yakın olması yerel bağımsızlık varsayımının karşılandığını göstermektedir (Hambleton, 1991; McDonald, 1981; Akt: Bulut, 2015). Bu noktadan hareketle bu çalışmada düşük ve yüksek yetenek gruplarındaki bireylerin maddelere verdikleri cevaplar üzerinden elde edilen maddeler arası korelasyonlar elde edilmiştir. Elde edilen sonuçlara göre her iki yetenek grubunda da maddeler arası korelasyonlar çok düşük çıkmıştır. Böylece birleşik test için yerel bağımsızlık varsayımının karşılandığı sonucuna ulaşılabilir.

Son olarak TBPOMTK için monotonluk varsayımı test edilmiştir. Bu çalışmada monotonluk varsayımının test edilmesi için R 3.0.2. yazılımı için Van Der Ark (2007) tarafından geliştirilen kullanılan Mokken paketinden yararlanılmıştır. Monotonluk varsayımının sonuçlarını yorumlamaya geçmeden önce önemli sembol ve kısaltmaları açıklamakta yarar vardır. (#AÇ) her bir madde için aktif çift sayısını, (#GMİ) gizil monotonluğun ihlalinin miktarı, (#GMİ/#AÇ) her madde çifti için ortalama olarak monotonluk ihlalinin miktarını, (maxGMİ) monotonluk ihlalinin miktarının en büyük değerini, (TOP) toplam monotonluk ihlalinin miktarını, (TOP/#AÇ) her madde çifti için toplam monotonluk ihlalinin miktarını göstermektedir. Bütün bu değerler 0'dan anlamlı bir şekilde büyük olduğu takdirde gizil monotonluk varsayımı ihlal edilmiş olur (Van der Ark, 2007). Yorumlama için önemli olan bir diğer gösterge ise madde ölçeklenebilirlik katsayısı olan Hj'dir. Her bir maddeye ait olan Hj'lerin 0,30'dan küçük olması maddenin ayırtedicilik bakımından zayıf olduğunu gösterir. Hj madde ayırt edicilik katsayısı olarak yorumlanır (Sijtsma ve Molenaar, 2002). Elde edilen sonuçlar doğrultusunda Din Kültürü ve Ahlak Bilgisi, İngilizce ve Fen bilgisi alt testleri için ise bazı maddeler tarafından varsayımın karşılanmadığı görülmüştür. Türkçe ve Matematik alt testlerinde yer alan tüm maddelerin ise varsayımı karşıladığı tespit edilmiştir. Özellikle Matematik testinde yer alan bütün maddeler için (#AÇ), (#GMİ), (#GMİ/#AÇ), (maxGMİ), (TOP) ve (TOP/#AÇ) değerleri 0 çıkmıştır. Benzer şekilde matematik testinde yer alan bütün maddelere ait ölçeklenebilirlik katsayısı 0,30'un üzerindedir.

Elde edilen sonuçlardan hareketle TBPMTK ve TBPOMTK analizlerinde tek boyutluluk kanıtlarının daha güçlü olduğu belirlenen ve TBPOMTK için monotonluk varsayımını da karşılayan Matematik alt testinin sonuçlarından yararlanılmıştır. ÇBMTK için ise daha önce ifade edildiği gibi Din Kültürü ve Ahlak Bilgisi ve Fen Bilgisi testlerinden belli maddelerin seçilmesi ile oluşturulan iki faktörlü birleşik testin sonuçlarından yararlanılmıştır. Tek ve çok boyutlu veri analizi için kullanılan testlerin belirlenmesinden sonra, tek ve çok boyutlu MTK kapsamında hangi modellerin kullanılacağının belirlenmesi gerekir. TBPMTK için matematik testinin sonuçlarına 2 PLM ile 3 PLM uygulanmıştır. 2 PLM için kestirilen parametre sayısı 40 ve elde edilen -2 LL değeri 1419674.370'tir. 3 PLM için ise kestirilen parametre sayısı 60 ve elde edilen -2 LL değeri 1702461.230'dur. 20 serbestlik derecesinde elde edilen fark değeri ise 282786.86'dır. Elde edilen bu sonuç 20 (60-40) serbestlik derecesindeki χ^2 kritik değeri (31.410) ile kıyaslandığında, anlamlı çıkmıştır. Dolayısıyla, -2 LL değeri düşük olan 2PLM'nin 3PLM'ye göre anlamlı farklılık yarattığı; diğer bir ifadeyle 2PLM'nin veriye daha iyi uyum sağlayan model olduğu söylenebilir. Bununla birlikte Embretson ve Reise (2000), çoktan seçmeli test maddesi ile çalışıldığı durumlarda 3PLM'nin, kişilik verileri ile çalışıldığı durumlarda ise 1PLM veya 2PLM'den birinin kullanılmasını önerir. Ancak modellere ait -2LL değerleri de göz önünde bulundurularak TBPMTK modeli olarak 2PLM'nin uygulanmasına karar verilmiştir.

ÇBMTK için birleşik testin sonuçlarına genişletilmiş M2PLM ile M3PLM uygulanmıştır. M2PLM için kestirilen parametre sayısı 40 ve elde edilen -2 LL değeri 1214446.23'dur. M3PLM için ise kestirilen parametre sayısı 60 ve elde edilen -2 LL değeri 1214490.81'dir. 20 serbestlik derecesinde elde edilen fark değeri ise 44.58'dir. Elde edilen bu sonuç 20 (60-40) serbestlik derecesindeki $\chi 2$ kritik değeri (31.410) ile kıyaslandığında, anlamlı çıkmıştır. Bu durumda analizler için tercih edilen model -2 LL değeri daha küçük olan genişletilmiş M2PLM'dir.

Son olarak TBPOMTK için Andries van der Ark (2007) tarafından önerilen İkili Monotonluk Modeli (IMM) ve Monoton Homojenlik Modeli (MHM) arasından MHM'nin kullanılmasına karar verilmiştir. Bunun nedenli IMM tarafından açıklanabilen her veri setinin daha zayıf bir Modeli (MHM) tarafından açıklanabilmesidir (Andries van der Ark, 2007)

TBPOMTK analizlerinde R 3.0.2. yazılımı için Van Der Ark (2007) tarafından geliştirilen MOKKEN paketinden, tek ve çok boyutlu MTK analizinde ise Cai (2017) tarafından önerilen FlexMIRT 3.5 yazılımından yararlanılmıştır. TBPMTK parametre kestirimi 2 PLM'ye göre yapılmış, çok boyutlu MTK analizleri ise genişletilmiş 2 PLM'ye göre gerçekleştirilmiştir. Hem TBPMTK hem de ÇBMTK analizlerinde hata değerleri Cai'nin (2008) EM algoritmasının hata değerleri kestirilerek belirlenmiştir.

Bu çalışmada madde puanlarından değil test puanlarından yararlanılmıştır. Bu nedenle toplam test puanları için ortalama güçlük ve ayırtedicilik katsayıları kestirilmiştir. Her üç kurama ait modeller için de testin ortalama güçlüğü ve ayırtedicilik düzeyleri evrenden birer kez çekilen her bir örneklem için ayrı ayrı hesaplanmıştır. Her bir örneklem için kestirilen parametre ortalamalarının evren değerden ne kadar farklı olduğu incelenmiştir. Ancak bu fark istatistiksel olarak test edilmemiştir. Yapılan yorumlar sadece büyüklük küçüklük ilişkisi çerçevesinde yapılmıştır.

Çalışmanın odağında olan parametre değişmezliğinin incelenmesi amacıyla, parametre değişmezliğinin göstergesi olarak standart hata ortalamalarından (SHO) yararlanılmıştır (Koğar, 2014; Sünbül, 2011). Ancak bulgular yorumlanırken parametre ortalamalarının kestirimlerine ait SHO'ların yanı sıra parametre ortalamalarının da öneklem büyüklüğünden nasıl etkilendiği incelenmiştir. Bu amaçla her bir örneklem için kestirilen parametre ortalamalarına ait SHO'ların evren değerden nekadar farklı olduğu incelenmiştir. Tıpkı parametre ortalamalarında olduğu gibi bu fark istatistiksel olarak test edilmemiştir. Araştırmanın amacı betimleme olduğundan elde edilen bulgular sadece büyüklük-küçüklük ilişkisi çerçevesinde yorumlanmıştır.

Çalışmaya ait bulguların elde edilmesi genel olarak iki başlık altında toplanabilir. Bunlardan ilki tek boyutlu olan Matematik testi tek boyutlu parametrik ve parametrik olmayan MTK altında modellenmiş, bu modellemeye göre madde parametreleri kestirilmiştir. İkinci boyutta ise iki boyutlu olan birleşik testin tek boyutlu olduğu kabulünden (iki boyutluluğun ihmalinden) yola çıkılarak tek boyutlu parametrik ve parametrik olmayan MTK'ye göre tek boyutlu bir model altında analiz

edilmiştir. Bunlara ek olarak iki boyutlu olan birleşik test doğasına uygun olarak ÇBMTK'ye göre modellenerek analiz edilmiş ve bu analizlerin sonucunda madde parametreleri kestirilmiştir.

Araştırmanın İç ve Dış Geçerliği

Araştırmanın iç geçerliği bağımlı değişkendeki değişimlerin, bağımsız değişkenlerle açıklanma derecesi ile ilgilidir. Bu çalışmada madde ve yetenek kestirimleri ile bu kestirimlere ait güvenirlik düzeylerindeki değişim, örneklem büyüklüğü ve MTK'nin farklı uygulamaları tarafından açıklanabildiği için araştırmanın iç geçerliği sağlanmıştır (Fraenkel & Wallen, 2006).

Araştırmanın dış geçerliği bulguların genellenebilirlik derecesi ile ilgilidir. Bu çalışmadan elde edilen bulguların genellenebilirliği kullanılan örneklem büyüklükleri, MTK uygulamaları ve kullanılan testlerin konu alanı ile sınırlı olduğu için araştırmanın dış geçerliğini bu çerçevede incelemek gerekir. Dolayısıyla kullanılan örneklem büyüklükleri, kullanılan MTK uygulamaları ve testlere ait konu alanı çerçevesinde araştırma sonuçlarının genellenebileceği düşünülmektedir (Fraenkel ve Wallen, 2006).

BULGULAR

Birinci Alt Problemin Çözümüne İlişkin Bulgular

Birinci alt problemin çözümü için TBPMTK ve TBPOMTK'ye göre kestirilen matematik testi'ne ait madde parametreleri ve standart hata ortalamaları (SHO) Tablo 1'de görülmektedir.

Tablo 1. BPMTK ve TBPOMTK'ye Göre Kestirilen Matematik Testi'ne Ait Madde Parametreleri ve Standart Hata Ortalamaları

Örneklem Büyüklüğü	TBPOMTK				TBPMTK			
	H	H_{SHO}	p	p_{SHO}	a	a_{SHO}	b	b_{SHO}
50	0.32	0.09	0.48	0.109	1.07	0.32	0.09	0.37
100	0.39	0.07	0.45	0.109	1.31	0.28	0.24	0.23
200	0.33	0.05	0.43	0.107	1.32	0.22	0.34	0.18
500	0.33	0.03	0.43	0.107	1.41	0.15	0.34	0.10
1000	0.33	0.02	0.43	0.107	1.48	0.11	0.30	0.07
5000	0.33	0.01	0.43	0.107	1.50	0.05	0.29	0.03
Evren	0.33	0.00	0.43	0.107	1.52	0.01	0.30	0.01

TBPOMTK'ye göre kestirilen tek boyutlu teste ait madde parametrelerinin ilki ayırtediciliğin göstergesi olan ortalama H parametreleridir. Tabloda verilen H ortalamaları incelendiğinde evren değerin 0.33 olduğu görülmektedir. Evren değere göreli olarak en uzak H ortalaması 100 örneklem büyüklüğünden kestirilmiştir (H=0.39). Evren değere en uzak H ortalaması bile evren değerden çok farklı değildir. Bu nedenle örneklemlerden kestirilen H ortalamalarının büyüklük olarak evren değere yakın olduğu sonucuna ulaşılabilir. Bununla birlikte 200 örneklem büyüklüğünden itibaren ise evren değeri yansıtacak düzeyde kararlı bir yapıya sahip H'lerin kestirildiği tabloda görülmektedir (H=0.33). H ortalamasına ait SHO'nun evren değeri 0'a çok yakındır. Evren değeri en az yansıtan örneklem büyüklüğünün en küçük örneklem olan 50 olduğu (H_{SHO}=0.09) ve örneklem büyüklüğü arttıkça H parametresi ortalamasına ait SHO'ların evren değerine yaklaştığı yine tabloda görülmektedir. Bunun yanı sıra örneklem büyüklüğü kaç olursa olsun SHO'lar büyüklük olarak evren değere yakın büyüklükte kestirilmiştir. Elde edilen bu sonuçlara göre evren değere yakın SHO'lar ile ortalama H parametresi kestirmek için büyük örneklemlerin kullanılmasının zorunlu olmadığı sonucuna ulaşılabilir.

TBPOMTK'ye göre kestirilen tek boyutlu teste ait madde parametrelerinin bir diğeri ise güçlük göstergesi olan ortalama p parametreleridir. Evren değerine en uzak p ortalaması en küçük örneklemde kestirilmiştir (p=0.48). Ancak Tablo 1'de görüldüğü gibi evren değere en uzak p ortalaması bile evren değerden çok farklı değildir.

Ortalama p parametresine ait SHO'nun evren değeri ise 0.107 olarak kestirilmiştir. Evren değere göreli olarak en uzak SHO ise 50 ve 100 örneklem büyüklüklerinden kestirilmiştir (H_{SHO} =0.109). SHO 200 örneklem büyüklüğünde ise en düşük değerini almıştır (p_{SHO} =0.107) ve bu değerin evren değere eşit olduğu söylenebilir. Tıpkı H ortalamalarına ait SHO'larda olduğu gibi örneklemlerden kestirilen p ortalamalarına ait SHO'lar ile evren değeri arasında büyük farklılıkların olmadığı belirlenmiştir. Bununla birlikte 200 örneklem büyüklüğü ile evren arasındaki örneklemlerde de SHO büyüklüklüklerinde bir değişim olmamıştır. Sonuç olarak TBPOMTK'ye göre kestirilen madde parametresi ortalamaları ve madde parametresi ortalamalarına ait SHO'ların 200 örneklem büyüklüğünden itibaren evreni yansıtacak düzeyde kararlı bir yapıya sahip olduğu görülmüştür. Buna göre TBPOMTK için 200 örneklem büyüklüğü itibariyle parametre değişmezliğinin sağlandığı sonucuna ulaşılmıştır.

TBPMTK'ye göre kestirilen tek boyutlu teste ait madde parametrelerinin ilki ayırtediciliğin göstergesi olan ortalama a parametreleridir. Örneklemlerden kestirilen a parametresi ortalamasının örneklem büyüklüğüne bağlı olarak 1.07 ile 1.52 arasında değerler aldığı Tablo 1'de görülmektedir. Ortalama a parametresinin evren değeri 1.52 olarak kestirilmiştir. Ortalama a parametresi evren değerine en uzak değerini en küçük örneklem büyüklüğü olan 50'de almıştır ve örneklem büyüklüğü arttıkça a parametresi ortalamasına ait değerlerin de artma eğiliminde olduğu ve evren değerine yaklaştığı yine tabloda görülmektedir. 1000 örneklem büyüklüğünden itibaren a parametresi ortalamaları evren değerine çok yakın büyüklükte kestirilmiştir.

TBPMTK için a parametresi ortalamasına ait SHO'nun evren değeri ise 0.01 olarak kestirilmiştir. Evren değerine en uzak SHO'nun 50 örneklem büyüklüğünden kestirildiği tabloda görülmektedir (a_{SHO} =0.32). Örneklem büyüklüğünün artmasıyla a parametresi ortalamasına ait SHO azalma eğilimi göstererek evren değere yaklaşmıştır. Tabloda görüldüğü gibi 500 örneklem büyüklüğü itibariyle evren değere çok daha yakın bir SHO ile a parametresinin kestirildiği sonucuna ulaşılabilir.

TBPMTK'ye göre kestirilen tek boyutlu teste ait madde parametrelerinin bir diğeri ise güçlük göstergesi olan ortalama b parametresidir. Ortalama b parametresi ait evren değer 0.30 olarak kestirilmiştir. Evren değerine en uzak b parametresi ortalaması 50 örneklem büyüklüğünden elde edilmiştir (b=0.09). Evren değeri en iyi yansıtan örneklem büyüklüğü ise 1000'dir. Elde edilen sonuçlara göre b parametresi ortalamasının, örneklem büyüklüğünün değişmesi ile düzenli bir artma ya da azalma eğiliminde olduğunu söylemek mümkün değildir. Ancak 1000 örneklem büyüklüğü itibariyle b parametresi ortalamasına ait değişimlerin daha az olduğu söylenebilir. Bu durumda 1000 örneklem büyüklüğü itibari ile b parametresi ortalamalarının evrene yakın büyüklükte olduğu sonucuna ulaşılabilir.

TBPMTK için b parametresi ortalamasına ait SHO'nun evren değeri ise 0.01 olarak kestirilmiştir. Evren değerine en uzak SHO en küçük örneklem büyüklüğü olan 50 üzerinden kestirilmiştir (b_{SHO} = 0.37). Ortalama b parametresi ait SHO'nun örneklem büyüklüğünün artmasına bağlı olarak istikrarlı bir şekilde evren değerine yaklaştığı Tablo 1'de görülmektedir. Tabloda görüldüğü gibi 500 örneklem büyüklüğünden itibaren de evren değerine daha yakın büyüklükte bir SHO ile b parametresinin kestirildiği sonucuna ulaşılabilir.

TBPMTK üzerinden elde edilen madde parametresi kestirimlerinin incelenmesinin sonucunda, a ve b parametreleri ile a ve b parametrelerine ait SHO'lar sürekli bir değişim gösterdiği için parametre değişmezliğinin sağlanamadığı sonucuna ulaşılmıştır.

İkinci Alt Problemin Çözümüne İlişkin Bulgular

İkinci alt problemin çözümü için birleşik testin sonuçlarından ÇBMTK ile TBPMTK'ye göre kestirilen birleşik test ait madde parametreleri ve standart hata ortalamaları (SHO) Tablo 2'de görülmektedir.

Tablo 2. Tek ve Çok Boyutlu MTK'ye Göre Kestirilen Birleşik Teste Ait Madde Parametreleri ve Standart Hata Ortalamaları

Örneklem Büyüklüğü	ÇBMTK						TBPMTK			
	a_1	$a_{1\ SHO}$	a_2	$a_{2\ SHO}$	d	d_{SHO}	a	a_{SHO}	b	b_{SHO}
50	0.98	1.47	0.51	0.75	1.12	0.99	1.23	0.54	-0.82	0.48
100	0.60	0.70	0.44	0.45	0.77	0.53	1.24	0.75	-0.74	0.47
200	0.68	0.57	0.42	0.30	1.02	0.43	1.48	0.52	-0.79	0.36
500	0.68	0.43	0.40	0.20	0.87	0.36	1.68	0.69	-0.55	0.19
1000	0.67	0.25	0.41	0.14	0.84	0.20	1.76	0.16	-0.48	0.09
5000	0.67	0.11	0.42	0.06	0.88	0.08	1.94	0.07	-0.49	0.03
Evren	0.41	0.01	0.43	0.01	-0.18	0.01	1.52	0.01	0.30	0.01

ÇBMTK'ye göre kestirilen iki boyutlu teste ait madde parametrelerinin ilki ayırtediciliğin göstergesi olan ortalama a_1 ve a_2 parametreleridir. ÇBMTK'den kestirilen a_1 ortalamasının evren değeri 0.41 olarak kestirilmiştir. Ortalama a_1 için evren değere en uzak kestirim en küçük örneklem büyüklüğü olan 50 üzerinden elde edilmiştir. a_1 ortalamasının evren değerini iyi yansıtan bir örneklem büyüklüğü ise bulunmamaktadır. Ortalama a_2'ye ait evren değeri ise 0.43 olarak kestirilmiştir. Ortalaa a_2'nin evren değerine en uzak kestirim en küçük örneklem büyüklüğünden elde edilmiş ve 100 örneklem büyüklüğünden itibaren ise evren değerine yakın büyüklükte a_2 ortalaması kestirimleri elde edilmiştir. Elde edilen sonuçlara göre a_1 ve a_2 ortalamalarının, örneklem büyüklüğünün değişimine bağlı olarak düzenli bir artma ya da azalma eğiliminin olduğunu söylemek mümkün değildir.

ÇBMTK için ayırtedicilik parametrelerinin ortalamasına ait SHO'nun evren değerinin ise 0,01 olarak kestirildiği tabloda görülmektedir. Evren değerine en uzak SHO kestirimi en küçük örneklem büyüklüğü üzerinden elde edilmiştir ($a_{1\ SHO}$=1.47; $a_{2\ SHO}$=0.75). Örneklem büyüklüğünün artmasıyla her iki ayırtedicilik parametresi ortalamasına ait SHO da azalma eğilimi göstermiştir ve evren değere yaklaşmıştır a_1 ortalamasının 5000 örneklemden itibaren ve a_2 ortalamasının 1000 örneklemden itibaren evren değere yakın büyüklükte bir SHO ile kestirildiği Tablo 2'de görülmektedir. Bununla birlikte her iki ayırtedicilik parametresi ortalamasına ait SHO da örneklemlerden evrene doğru sürekli bir değişim göstermektedir.

Sonuç olarak ÇBMTK'den kestirilen ayırtedicilik parametrelerine ait ortalama değerleri de SHO'ları da sürekli bir değişim gösterdikleri için parametre değişmezliğinin sağlanamadığı sonucuna ulaşılmıştır.

ÇBMTK'den kestirilen d parametresi ortalamasının evren değeri (d =-0.18) olarak kestirilmiştir. Evren değerine en uzak d parametresi ortalaması en küçük örneklem büyüklüğünden elde edilmiştir (d=1.12). Örneklemlerden elde edilen d parametresi ortalamaları incelendiğinde evren değerden farklı oldukları görülmektedir. Elde edilen sonuçlara göre d parametresi ortalamasının örneklem büyüklüğünün değişimine bağlı olarak düzenli bir artma ya da azalma eğiliminin olmadığı da görülmektedir. Ancak 500 ve 5000 örneklem büyüklükleri arasındaki değişimin diğer örneklem büyüklüklerine göre daha az olduğu Tablo 2'de görülmektedir.

ÇBMTK için d parametresinin ortalamasına ait SHO'nun evren değeri ise 0.01 olarak kestirilmiştir. Evren değerine en uzak SHO kestiriminin en küçük örneklem büyüklüğünden elde edildiği ve örneklem büyüklüğü arttıkça SHO kestirimlerinin de evren değerine yaklaştığı tabloda görülmektedir.

d parametresi ortalamasına ait SHO'ların örneklemler boyunca farklı değerler aldığı yine Tablo 2'de görülmektedir.

Tıpkı ayrıtedicilik parametrelerinin ortalamalarında olduğu gibi d parametresi ortalaması ile d parametresi ortalamasına ait SHO da örneklemlerden evrene doğru sürekli bir değişim gösterdiği için parametre değişmezliğinin sağlanamadığı sonucuna ulaşılmıştır.

TBPMTK'ye göre kestirilen iki boyutlu teste ait madde parametrelerinin ilki ayırtediciliğin göstergesi olan ortalama a parametreleridir. Örneklemlerden kestirilen a parametresi ortalamasının örneklem büyüklüğüne bağlı olarak 1.23 ile 1.94 arasında değerler aldığı Tablo 2'de görülmektedir. a parametresi ortalamasının evren değeri 1.52 olup evren değerini en iyi yansıtan örneklem büyüklüğü 200'dür. Daha genel bir ifade ile bu çalışmada kullanılan orta büyüklükteki örneklemlerde evren değere daha yakın kestirimler elde edilirken, uçtaki (en büyük ve en küçük) örneklemlerde evren değere daha uzak kestirimler elde edilmiştir. Ayrıca a parametresi ortalamasının evren değerinin tek boyutlu testten elde edilen evren değer ile aynı olduğunu da belirtmek gerekir. Buradaki karşılaştırmayı anlamlı hale getiren nokta, karşılaştırılan değerlerin aynı evren üzerinden fakat iki farklı boyutluluğa sahip yapılardan elde edilmesidir. Bu noktadan hareketle ulaşılan sonuç tek boyutluluk varsayımının sağlandığı ve sağlanmadığı durumlarda bir başka ifadeyle tek ve çok boyutlu testlerden kestirilen a parametresine ait evren değerlerin ortalamasının değişmediğidir.

TBPMTK için a parametresi ortalamasına ait SHO'nun evren değeri 0.01 olarak kestirilmiştir. Evren değere büyüklük olarak en uzak SHO 100 örneklem büyüklüğünden elde edilmiştir (a_{SHO}=0.75). a parametresi ortalamasına ait SHO'nun örneklem büyüklüğüne bağlı olarak düzenli bir artma ya da azalma eğilimi gösterdiği söylenemez. Ortalama a parametresine ait SHO ardışık örneklem büyüklüklerinin bazılarında evren değere daha uzak bir değer alırken bazılarında daha yakın bir değer almıştır. Çok boyutlu veri yapısının tek boyutlu bir model altında modellenmesi bu düzensizliğin başlıca sebebidir. Ortalama a parametresineait SHO'ların 1000 öneklem büyüklüğüne kadar evreni yansıtmayan değerler aldığı yine tabloda görülmektedir. 500 örneklem büyüklüğünden itibaren ise istikrarlı bir şekilde azalarak evren değere yaklaşmıştır. Bunun yanı sıra a parametresi ortalamasına ait SHO'lar en küçük örneklem büyüklüğünden en büyük örneklem büyüklüğüne kadar sürekli bir değişim göstermektedir.

TBPMTK'ye göre kestirilen iki boyutlu teste ait madde parametrelerinin bir diğeri ise güçlüğün göstergesi olan ortalama b parametreleridir. b parametresi ortalamasının evren değeri 0.30 olarak kestirilmiştir. Evren değerine büyüklük olarak en uzak kestirim örneklem büyüklüğünün 50 olduğu durumda elde edilmiştir (b=-0.82). Bununla birlikte b parametresi ortalamasının evren değerini iyi yansıtan bir örneklem büyüklüğü bulunmamaktadır. a parametresi için ifade edilen "tek boyutluluk varsayımının sağlandığı ve sağlanmadığı durumlarda bir başka ifadeyle tek ve çok boyutlu testlerden kestirilen a parametresine ait evren değerlerin değişmediği" sonucu b parametresi için de geçerlidir. Bir başka ifadeyle kullanılan veri seti ister tek boyutlu olsun ister çok boyutlu olsun kestirilen b parametrelerine ait evren değerler birbiri ile aynıdır. Elde edilen sonuçlara göre b parametresi ortalamasının örneklem büyüklüğünün değişimine bağlı olarak net bir artma ya da azalma eğiliminin olmadığı görülmektedir.

TBPMTK için b parametresi ortalamasına ait SHO'nun evren değeri 0.01 olup, evren değerine büyüklük olarak en uzak b parametresi ortalamasına ait SHO en küçük örneklem büyüklüğü olan 50 üzerinden kestirilmiştir (b_{SHO}= 0.48). Örneklem büyüklüğü arttıkça kestirilen SHO'lar da evren değerine yaklaşmıştır. 500 örneklem büyüklüğü itibariyle evren değerine çok yakın büyüklükte bir SHO ile b parametresinin kestirildiği de Tablo 2'de görülmektedir SHO (b_{SHO}=0.19). Tıpkı a parametresi ortalamasında olduğu gibi b parametresi ortalamasına ait SHO'lar da farklı örneklem büyüklüklerinde farklı değerler almıştır.

Sonuç olarak, TBPMTK'den kestirilen madde parametrelerinin ve madde parametrelerine ait SHO'ların incelenmesinin sonucunda, a ve b parametreleri ile a ve b parametrelerine ait SHO'lar sürekli bir değişim gösterdiği için parametre değişmezliğinin sağlanamadığı sonucuna ulaşılmıştır.

Üçüncü Alt Problemin Çözümüne İlişkin Bulgular

Üçüncü alt problemin çözümü için birleşik testin sonuçlarından TBPMTK ve TBPOMTK'ye göre kestirilen birleşik teste ait madde parametreleri ve standart hata ortalamaları (SHO) Tablo 3'te görülmektedir.

Tablo 3. TBPMTK ve TBPOMTK'ye Göre Kestirilen Birleşik Teste Ait Madde Parametreleri ve Standart Hata Ortalamaları

Örneklem Büyüklüğü	TBPMTK				TBPOMTK			
	a	a_{SHO}	b	b_{SHO}	H	H_{SHO}	p	p_{SHO}
50	1.23	0.54	-0.82	0.48	0.46	0.10	0.65	0.097
100	1.24	0.75	-0.74	0.47	0.38	0.07	0.65	0.096
200	1.48	0.52	-0.79	0.36	0.39	0.05	0.68	0.096
500	1.68	0.69	-0.55	0.19	0.38	0.03	0.65	0.096
1000	1.76	0.16	-0.48	0.09	0.40	0.02	0.65	0.096
5000	1.94	0.07	-0.49	0.03	0.40	0.01	0.66	0.095
Evren	1.52	0.01	0.30	0.01	0.33	0.00	0.43	0.107

Birleşik testin TBPMTK'ye göre analiz edilmesi ile elde edilen sonuçlar ikinci alt problemin çözümünde detaylı bir şekilde ifade edilmiştir. Birleşik testin TBPOMTK'ye göre analiz edilmesiyle elde edilen ilk parametre H parametresidir.

H ortalamasının evren değerinin 0.33 olduğu tabloda görülmektedir (H=0.33). Evren değere en uzak Hi ortalaması 50 örneklem büyüklüğünden kestirilmiştir (H=0.46). Bu durumda tek boyutluluk varsayımının sağlandığı ve sağlanmadığı durumlarda bir başka ifadeyle tek ve çok boyutlu testlerden kestirilen H parametresine ait evren değerlerin değişmediği sonucuna ulaşılabilir. H ortalaması açısından evreni en iyi yansıtan örneklem büyüklükleri 100 ve 500 örneklem büyüklükleridir.

H ortalamasına ait SHO için 0'a çok yakın büyüklükte bir evren değer kestirilmiştir (H_{SHO}= 0.00). Evren değere büyüklük olarak en uzak SHO en küçük örneklem büyüklüğü olan 50 üzerinden kestirilmiştir (H_{SHO}=0.10). H ortalamasına ait SHO'ların örneklem büyüklüğü arttıkça istikrarlı bir şekilde azalarak evren değerine yaklaştığı görülmektedir. Bunun yanı sıra örneklemlerden elde edilen SHO'ların evren değerinden büyük farklılıklar göstermediği belirlenmiştir. Bu durumda çok boyutlu bir veri üzerinden küçük örneklem büyüklükleri ile çalışıldığı zaman, evren değere yakın SHO'lar ile ayırtedicilik parametresinin kestirilmesinin mümkün olduğu söylenebilir. Ayrıca tek ve çok boyutlu testler TBPOMTK'ye göre analiz edildiğinde elde edilen H parametrelerine ait SHO'ların birbirlerine çok yakın değerler aldıkları görülmüştür. Bu durumda tek boyutluluğun ihlal edildiği durumda H ortalamalarının benzer SHO değerleri ile kestirildiği sonucuna ulaşılabilir. Bununla birlikte SHO'lar örneklemler boyunca büyüklük olarak sürekli bir değişim göstermektedir.

Tablo 3'te görüldüğü gibi H ortalamaları ve SHO'ları örneklemler boyunca birbirleri ile farklılık gösterdikleri için H ortalaması için de değişmezliğin sağlanamadığı sonucuna ulaşılabilir. Tek boyutluluğun sağlandığı durumda ise küçük bir örneklem büyüklüğünden itibaren değişmezliğin sağlandığı birinci alt problemin bulgularında ifade edilmiştir. Tek boyutlu modellerin kullanılması için sağlanması gereken en önemli varsayım olan tekboyutluluğun ihlali, TBPOMTK sonuçlarını değişmezlik açısından etkilemiştir.

TBPOMTK'ye göre kestirilen iki boyutlu teste ait madde parametrelerinin bir diğeri ise güçlüğün göstergesi olan ortalama p parametreleridir. p ortalamalarına göre testin örneklemlerde göreli olarak daha kolay olduğu da Tablo 3'te görülmektedir. Bununla birlikte p ortalamalarının evren değeri ile büyüklük olarak evrene en uzak değeri arasında büyük farklılıkların olmadığı da görülmektedir. p ortalamasının evren değeri 0.43 olarak kestirilmiştir (p=0.43). Büyüklük olarak evrene en uzak p ortalaması ise 200 örneklem büyüklüğünden kestirilmiştir. Tıpkı H ortalamasının evren değerinde

olduğu gibi olduğu gibi p ortalamasına ait evren değerinin de tek boyutlu testin TBPOMTK'ye göre analiz edildiğinde kestirilen evren değer ile aynı olduğu görülmüştür. Bu durumda yine tek boyutluluk varsayımının sağlandığı ve sağlanmadığı durumlarda bir başka ifadeyle tek ve çok boyutlu testlerden elde edilen ortalama p parametresine ait evren değerlerin değişmediği sonucuna ulaşılabilir.

Tablo 3'te görüldüğü gibi p ortalamalarına ait SHO'nun evren değeri 0.107 olarak elde edilmiştir (p_{SHO}=0.107). Evren değere en uzak SHO değeri ise 5000 örneklem üzerinden kestirilmiştir (p_{SHO}=0.095). Tek boyutlu test TBPOMTK'ye göre analiz edildiğinde evrenden elde edilen p ortalamalarının aynı SHO ile kestirilmesi dikkat çeken bir bulgudur. Buna göre tek boyutluluğun sağlandığı ve ihlal edildiği durumda p ortalamalarına ait SHO'ların aynı olduğu sonucuna ulaşılabilir. p ortalamalarına ait SHO'ların evren değeri ile evrene en uzak değeri arasında büyük farklılıkların olmadığı Tablo 3'te görülmektedir. Bu nedenle örneklem büyüklüğünün değişmesinin p ortalamalarının kestirildiği SHO'lar üzerinde büyük farklılıklara neden olmadığı söylenebilir. Bu durumda tek boyutluluğun sağlanamadığı ve büyük örneklemlere ulaşılamadığı takdirde evren değere yakın büyüklükteki SHO'lar ile ortalama p parametresinin kestirilmesinin mümkün olduğu sonucuna ulaşılabilir. p ortalamasına ait SHO'lar en büyük ve en küçük örneklem büyüklüklerinde farklı değerler almış, diğer örneklem büyüklüklerinde ise değişmemiştir.

Ayrıca ortalama p değerleri ve SHO'ları en büyük ve en küçük örneklem büyüklüklerinde farklı değerler aldığı için p ortalaması için de değişmezliğin sağlanamadığı sonucuna ulaşılmıştır. Tek boyutluluğun sağlandığı durumda ise küçük bir örneklem büyüklüğünden itibaren değişmezliğin sağlandığı birinci alt problemin bulgularında ifade edilmiştir. H ortalaması için de benzer sonuçlar ortaya konmuştur.

TBPMTK ile TBPOMTK'den elde edilen sonuçlar karşılaştırılacak olursa, H ortalamalarının a parametresi ortalamalarına göre küçük örneklem büyüklüklerinde bile evren değerine yakın SHO'lar ile kestirildiği Tablo 3'te görülmektedir. Tablo 3'ten elde edilen sonuçlara göre, a parametresi ortalamasını evren değerine yakın bir SHO ile kestirebilmek için en az 1000 hatta 5000 örneklem büyüklükleri ile çalışmak gerekmektedir. Daha küçük bir örneklemden evren değerine yakın bir SHO ile ortalama ayırtedicilik parametreleri kestirilebildiği için TBPOMTK'nin TBPMTK'ye göre daha avantajlı olduğu sonucuna ulaşılabilir. Tek boyutlu test parametrik ve parametrik olmayan MTK'ye göre analiz edildiğinde de benzer bir sonuca ulaşılmıştır. O halde tek boyutluluk varsayımı karşılansa da karşılanmasa da evren değerine yakın büyüklükte bir SHO ile ortalama ayırtedicilik parametresi kestirmek için, TBPMTK için büyük örneklemlerin kullanılması gerekmektedir. TBPOMTK için ise böyle bir sınırlama yoktur. Benzer yorumlar parametrik ve parametrik olmayan MTK'den elde edilen ortalama güçlük parametreleri için de geçerlidir. Evren değerine yakın büyüklükte bir SHO ile ortalama b parametresinin kestirilmesi için en az 1000 örneklem ile çalışmak gerekmektedir. Ancak evren değerine yakın bir SHO ile ortalama p parametresinin kestirilmesi için daha küçük örneklem büyüklükleri ile de çalışılabileceği Tablo 3'te görülmektedir.

SONUÇLAR ve TARTIŞMA

Birinci alt problemin bulgularından elde edilen sonuçlar doğrultusunda tek boyutlu testin TBPOMTK'ye göre analiz edilmesiyle 200 örneklem büyüklüğü itibariyle madde parametresi değişmezliğini sağlandığı görülmüştür. Bu durumda Büyük örneklem büyüklüklerine ulaşılamadığı durumda ortalama madde parametresi kestirimi amacıyla küçük örneklemlerden elde edilen veriler için TBPOMTK uygulamasından yararlanılabildiği sonucuna ulaşılmıştır. Elde edilen bu sonuç MTK'nin özellikle okul uygulamalarında eğitimin farklı kademelerinde uygulanan sınavlara ait madde parametresi kestirimlerine imkân sağlaması açısından büyük bir önem taşımaktadır.

Tek boyutlu testin TBPMTK'ye göre analiz edilmesiyle TBPMTK için örneklem büyüklüğü arttıkça a parametresi ortalamasına ait SHO'nun azaldığı ve evren değerine yaklaştığı görülmüştür. Thissen ve Wainer (1982) parametre kestirimi için 10.000 ve daha fazla örnekleme ihtiyaç olduğunu, Goldman ve Raju (1986) ise a parametresinin doğru kestirimi için en az 1000 kişilik örnekleme ihtiyaç duyulduğunu ifade etmişlerdir. Örneklem büyüklüğü 5000'i geçtiği halde a parametresine ait ortalama

değerin ve a parametresi ortalamasına ait standart hata ortalamasının değişmeye devam etmesi, Thissen ve Wainer (1982) bulgularını destekler niteliktedir.

Tek boyutlu testin TBPMTK'ye göre analiz edilmesiyle elde edilen b parametresi ortalamasına ait SHO örneklem büyüklüğü arttıkça azalma eğilimi göstermiş ve evren değerine yaklaşmıştır. b parametresi ortalaması ise örneklem büyüklüğüne bağlı olarak düzenli bir artma ya da azalma eğilimi göstermemiştir. Sünbül (2011) bu durumu "örneklem büyüklüğünün b parametresi üzerindeki önemsizliği" şeklinde ifade etmiştir. Bununla birlikte Hulin, Lissak ve Drasgow (1982) 2 PLM için 1000'den daha büyük örneklemlerde kestirilen parametrelerin önemli değişikliklerin olmadığını ortaya koymuştur. Bu nedenle elde edilen bulgular Hulin, Lissak ve Drasgow (1982)'nin bulgularıyla tutarlılık göstermektedir.

Sonuç olarak küçük örneklem büyüklüklerinden elde edilen madde parametrelerinin evreni yansıtacak düzeyde kararlı olması, bir başka ifadeyle küçük bir örneklem büyüklüğünden itibaren parametre değişmezliğinin sağlanması TBPOMTK'nin TBPMTK'ye göre önemli avantajlara sahip olduğunun bir kanıtıdır. Bu nedenle özellikle okul uygulamalarında TBPOMTK uygulamaları TBPMTK'ye tercih edilebilir.

İkinci alt problemin bulgularından elde edilen sonuçlar doğrultusunda birleşik testin ÇBMTK'ye göre analiz edilmesiyle elde edilen a_1 ve a_2 parametreleri ortalamasının, örneklem büyüklüğüne bağlı olarak düzenli bir artma ya da azalma eğilimi göstermediği sonucuna ulaşılmıştır. Ackermann (2005) ise örneklem büyüklüğü arttıkça maddelerin ayırtedicilik gücünün arttığı sonucuna ulaşmıştır. Bu açıdan çalışmadan elde edilen sonuç, Ackermann (2005)'in elde ettiği sonuç ile farklılık göstermektedir.

ÇBMTK analizleri sonuçlarına göre d parametresi ortalamasının örneklem büyüklüğünün değişimine bağlı olarak net bir artma ya da azalma eğiliminin olmadığı görülmektedir. d parametresi ortalaması ile d parametresi ortalamasına ait SHO da örneklemlerden evrene doğru sürekli bir değişim gösterdiği için parametre değişmezliğinin sağlanamadığı sonucuna ulaşılmıştır.

Birleşik test ÇBMTK'ye göre analiz edildiğinde her bir boyut için elde edilen a parametrelerinin ortalamasının ikisi de (0,1) aralığında değerler almıştır. TBPMTK'ye göre kestirilen ortalama ayırtedicilikparametreleri ise 1'in üzerinde değerler almıştır. Elde edilen bu sonuç Ansley ve Forsyth (1985)'in, iki kuramdan birbirine yakın büyüklükte a parametresi kestirimini elde ettiği çalışmasında ortaya koyduğu sonuçlardan farklılık göstermektedir. Bununla beraber çok boyutlu testin TBPMTK'ye göreanaliz edilmesinden elde edilen sonuçlara göre a ve b parametresi ortalamaları ve a ve b ortalamalarına ait SHO'lar bütün örneklemlerde farklı değerler aldığı için parametre değişmezliğinin sağlanamadığı sonucuna ulaşılmıştır.

Sonuç olarak çok boyutlu bir veri ister TBPMTK altında modellensin ister ÇBMTK altında modellensin evren değerine yakın bir büyüklükte SHO ile ortalama bir parametre kestirimi yapmak için en az 5000 örneklem büyüklüğünün kullanılması gerekir. 5000 ya da daha büyük örneklemlerin kullanılması durumunda kestirilen SHO evren değerine yaklaşmış olur. Ancak her ne kadar parametre ortalamalarına ait SHO evren değerine yaklaşsa da çok boyutlu bir verinin tek boyutlu modele göre analiz edilmesi ile elde edilen parametre ortalamaları ile çok boyutlu modele göre analiz edildiğinde elde edilen parametre ortalamaları örneklem büyüklüğüne bağlı olmaksızın birbirinden farklı çıkmıştır. Bir başka ifade ile parametre değşmezliği sağlanmamıştır. Elde edilen bu sonuç tek boyutluluk varsayımı sağlanmadığı zaman parametre kestirimlerine ait sonuçların farklılaştığını ifade eden Drasgow ve Parsons (1983)'ün bulgularıyla tutarlılık göstermektedir. Burada değişmezliğin sağlanamamasının ve parametre ortalamalarının örneklem büyüklüğüne bağlı olarak net bir artma ya da azalma göstermemesinin nedeni olarak bu çalışmada örneklem seçiminde replikasyon yapılmamış olması düşünülmektedir.

Üçüncü alt problemin bulgularından elde edilen sonuçlar doğrultusunda birleşik testin TBPOMTK'ye göre analiz edilmesiyle elde edilen H ortalamasına ait SHO'ların örneklem büyüklüğü arttıkça istikrarlı bir şekilde azalarak evren değerine yaklaştığı sonucuna ulaşılmıştır. Koğar (2014) çalışmasında örneklem büyüklüğü arttıkça H katsayılarına ait standart hatanın azaldığını ortaya koymuştur. Elde edilen bu sonuç Koğar (2014)'ün bulgularıyla tutarlılık göstermektedir. Bunun yanı

sıra örneklem büyüklüğü kaç olursa olsun SHO'lar büyüklük olarak evrene çok yakın değerler almıştır. Bu durumda tek boyutluluğun ihlal edilmesinin Hi ve p ortalamalarının kestirildiği SHO değerlerinin üzerinde önemli bir etkisinin olmadığı sonucuna ulaşılmıştır.

Tek boyutluluk varsayımının ihlal edilmesinin parametre kestirimleri üzerindeki en önemli etkisi, tek boyutluluk varsayımı karşılandığı durumda parametre değişmezliği sağlanırken, varsayım karşılanmadığında parametre değişmezliğinin de sağlanamaması olmuştur. Bu nedenle bu çalışmada Koğar (2014)'ün örneklem büyüklüğü arttıkça p değerleri için değişmezliğin sağlandığı bulgusuna ulaşılamamıştır. Koğar (2014)'ün elde ettiği bu sonuç, 1. alt problemden elde edilen bulgularda ifade edildiği gibi tek boyutlu veriye ait p ortalamaları için elde edilmiştir. Bir başka ifadeyle tek boyutlu veriye ait p ortalamaları için örneklem büyüklüğü arttıkça parametre değişmezliği sağlanmıştır.

TBPMTK ile TBPOMTK'den elde edilen sonuçlar karşılaştırıldığında TBPOMTK'den kestirilen parametre ortalamalarının evren değerine yakın bir SHO ile kestirilebildiği sonucuna ulaşılmıştır. TBPMTK'den elde edilen parametre ortalamalarının evren değerine yakın bir SHO ile kestirilebilmesi için ise en az 1000 örneklem büyüklüğü ile çalışmak gerekmektedir. Daha küçük bir örneklemden evren değerine yakın bir SHO ile madde parametrelerinin kestirilebilmesiTBPOMTK'nin TBPMTK'ye göre avantajlı olduğunun bir kanıtıdr. Bununla birlikte tek boyutluluk varsayımının karşılanıp karşılanmamasının bu sonucu değiştirmediği dikkat çeken bir bulgudur.

Öneriler

Bu çalışmada örneklem büyüklüğünün etkisini incelemek için 50, 100, 200, 500, 1000 ve 5000 örneklem büyüklükleri ile çalışılmıştır. Benzer bir çalışma farklı örneklem büyüklükleri ile de gerçekleştirilebilir. Bu çalışmada TBPMTK için 2PLM, TBPOMTK için MHM ve ÇBMTK için M2PLM kullanılmıştır. Sonuçları incelenen kuramlara ait olan farklı modeller ile benzer bir çalışma yapılabilir.

Bu çalışmada kullanılan modeller üzerinden kestirilen parametre değerleri tüm teste ait ortalama parametre değerleridir. Benzer bir çalışma bir testte yer alan maddelere ait parametre kestirimleri için de gerçekleştirilebilir. Bu çalışmada örneklem büyüklüğünün etkisini araştırmak amacıyla parametre kestirimlerine ait standart hata ortalaması kullanılmıştır. Başka bir çalışmada örneklem büyüklüğünün etkisini araştırmak için farklı göstergelerden de yararlanılabilir.

Bu çalışmada şans başarısının etkisi göz önünde bulundurulmamıştır. Şans başarısından arındırılmış gerçek puanlar üzerinden de başka bir çalışma yapılabilir. Bu çalışmada örneklem büyüklüğünün parametre değişmezliğinin üzerindeki etkisi araştırılmıştır. Farklı bağımsız değişkenler ile de benzer bir çalışma yapılabilir. Bu çalışmada örneklem seçiminde her bir örneklem büyüklüğü yalnızca bir kez seçilmiştir. Başka bir çalışmada replikasyon yoluyla örneklem seçerek örneklem büyüklüğünün parametre değişmezliği üzerindeki etkisi araştırılabilir.

KAYNAKÇA

Ackerman, T. A. (2005). Multidimensional item response theory modeling. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Multivariate applications book series. Contemporary psychometrics: A festschrift for Roderick P. McDonald* (p. 3–25). Lawrence Erlbaum Associates Publishers.

Andries van der Ark, L. (2007). Mokken Scale Analysis in R. *Journal of Statistical Software*, *20*(11).1-19.

Ansley, T.N. and Forsyth, R.A. (1985). An Examination of the characteristics of Unidimensional IRT estimates derived from two dimensional data. *Applied Psychological Measurement, 9*(1), 37-48.

Şengül Avşar A. (2018). Kategori sayısının psikometrik özellikler üzerine etkisinin mokken homojenlik modeli'ne göre incelenmesi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 9*(1), 49-63.

Şengül Avşar A. , Tavşancıl E. (2017). Examination of polytomous items' psychometric properties according to nonparametric item response theory models in different test conditions, *Kuram ve Uygulamada Egitim Bilimleri,* 17, 493-514.

Cai L. (2017). *flexMIRT Version 3.51: Flexible multilevel multidimensional item analysis and test scoring* (Computer software). Chapel Hill, NC: Vector Psychometric Group.

Cai, L. (2008). SEM of Another Flavour: Two new applications of the supplemented em algorithm. *British*

Journal of Mathematical and Statistical Psychology, 61, 309–329.

Çokluk, Ö., Şekercioğlu, G. ve Büyüköztürk, Ş. (2010). *Sosyal bilimler için çok değişkenli istatistik teknikleri.* Ankara: Pegem Akademi.

Drasgow, F. & Parsons,K. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement, 7*(2), 189-199.

Embretson, S. E., & Reise, S. P. (2000). *Multivariate applications books series. Item response theory for psychologists.* Lawrence Erlbaum Associates Publishers.

Fraenkel, J.R. & Wallen, N.E. (2006). *How to design and evaluate research in education* (Sixth edition). Boston: McGraw-Hill Pub.

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

Hambleton, R. K. & Swaminathan, H. (1985). *Item Response Theory: Principles and applications*. Boston: Academic Puslishers Group.

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*(3), 249-260.

Harman, H. H. (1962). *Modern Factor Analysis* (2. edition). University of Chicago Press, Chicago.

Kaptan, S. (1977). *Bilimsel Araştırma Teknikleri,* Ankara: Tekışık Matbaası ve Rehber Yayınevi.

Koğar H. (2018). Examining invariant item ordering using mokken scale analysis for polytomously scored items. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 9*(4), 312-325.

Koğar, H. (2014). *Madde tepki kuramının farklı uygulamalarından elde edilen parametrelerin ve model uyumlarının örneklem büyüklüğü ve test uzunluğu açısından karşılaştırılması.* (Doktora tezi, Hacettepe Üniversitesi, Eğitimde Ölçme ve Değerlendirme Anabilim Dalı, Ankara). http://tez2.yok.gov.tr/ adresinden edinilmiştir.

Meara, K., Robin, F. & Sireci, S.G. (2000). Using multidimensional scaling to assess the dimensionality of dichotomous item data. *Multivariate Behavioral Research, 35*(2), 229–259.

Mor-Dirlik, E. (2017). *Parametrik ve parametrik olmayan madde tepki kuramı modellerinden çeşitli faktörlere göre elde edilen madde ve yetenek kestirimlerinin karşılaştırılması*. (Doktora tezi, Ankara Üniversitesi, Eğitimde Ölçme ve Değerlendirme Anabilim Dalı, Ankara). http://tez2.yok.gov.tr/ adresinden edinilmiştir.

Price, L. R. (2017). *Psychometric methods: Theory and practice*. New York, NY: The Guilford Press.

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to Nonparametric Item Response Theory.* Thousand Oaks, CA: Sage.

Smits, I. A. M., Timmerman, M. E., & Meijer, R. R. (2012). Exploratory mokken scale analysis as a dimensionality assessment tool: why scalability does not ımply unidimensionality. *Applied Psychological Measurement, 36*(6), 516-539.

Sodano, S. M., and Tracey, T. J. G. (2011). A brief inventory of interpersonal problems- circumplex using non-parametric item response theory: introducing the IIP-C- IRT. *Journal of Personality Assessment, 93*(1), 62-75. doi:10.1080/00223891.2010.528482

Syu, J. J. (2013). *Applying person fit-in faking detection-the simulation and practice of non parametric item response theory*. (Doctoral Dissertation, National Chengchi University). Retrieved from http://nccur.lib.nccu.edu.tw/bitstream/140.119/58646/1/251501.pdf

Sünbül, Ö. (2011). *Çeşitli boyutluluk özelliklerine sahip yapılarda, madde parametrelerinin değişmezliğinin klasik test teorisi, tek boyutlu madde tepki kuramı ve çok boyutlu madde tepki kuramı çerçevesinde incelenmesi.* (Doktora tezi, Mersin Üniversitesi, Eğitimde Ölçme ve Değerlendirme Anabilim Dalı, Ankara). http://tez2.yok.gov.tr/ adresinden edinilmiştir.

Tabachnick, G. B. ve Fidell, S. L. (2001). *Using multivariate statistics* (4th Edition), Boston MA: Allyn&Bacon.

Thissen, D., & Wainer, H. (1982). Some standart errors in item response theory. *Psychometrika, 47*(4), 397-412.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2020; 11(2); 113-130

A Measurement Tool for Repeated Measurement of Assessment of University Students' Writing Skill: Development and Evaluation *

Ayfer SAYIN ** Nilüfer KAHRAMAN ***

Abstract
This study summarizes the development and the application of a four-week repeated assessment scale that was designed to measure the advanced writing skills of college students. The storyline was written by the first author and required respondents to write one or more sentences to the given developing storyline of the week. Each week's narrative was written to tell a part of a story that integrates into a single storyline over the weeks. For the application, each week, the respondents from a volunteered sample of 74 were asked 1) to write to a continuation to the developing storyline of the week (the instruction stated that this was the first or the second or the third or the last part of the storyline) and 2) to rate their overall mood that week (1 to 5, 5 indicating a great mood). Writings of the students (responses) were then coded by multiple raters with respect to three subskill components; namely expression, aesthetics, and creativity. The hypothesized tie between the fluctuations observed in the sense of well-being and the writing performance of the students over the weeks and whether and to what extent the creativity subcomponent was more subject to the influence of student's mood changes when compared to the clarity of expression or the aesthetics subskill. However, the results show that when the changes in writing performances of the whole group were examined instead of that of individuals over time, there were no significant differences to be found. It is recommended that it might be more useful than the classical one-shot assessment design.

Key Words: Longitudinal study, longitudinal measurement tool, expression skill, perception of aesthetics, creativity skill.

INTRODUCTION

With the aim of educational reforms for a high quality, future-oriented education responding to the needs of the society, the 2006 European Parliament and Council published their recommendations on the Key Competencies for Lifelong Learning. The eight key competencies that students are recommended to be equipped with are listed as follows: communicating in the mother tongue, communicating in a foreign language, mathematical, scientific and technological competence, digital competence, learning to learn, social and civic competence, sense of initiative and entrepreneurship, and cultural expression and awareness. In this report, it is stated that each competency is indeed highly associated with each other, and all the competencies intersect with critical thinking, problem solving, creative thinking, and establishing empathy (European Commission-EC, 2006). Similarly, the first of competence of Turkey Qualifications Framework is given as:

> Concepts, thoughts, ideas, emotions, and phenomena to express both orally and in writing and interpretation (listening, speaking, reading and writing); to interact in an appropriate and creative way linguistically in all social and cultural contexts such as

* This study was presented at the 6th International Congress on Measurement and Evaluation in Education and Psychology held in Kosovo on September 05-08, 2018.
** Assoc. Prof., Gazi University, Faculty of Gazi Education, Ankara-Turkey, ayfersayin@gazi.edu.tr, ORCID ID: 0000-0003-1357-5674
*** Prof., Gazi University, Faculty of Gazi Education, Ankara-Turkey, nkahraman@gazi.edu.tr, ORCID ID: 0000-0003-2523-0155

To cite this article:
Sayın, A., & Kahraman, N. (2020). A measurement tool for repeated measurement of assessment of university students' writing skill: Development and evaluation. *Journal of Measurement and Evaluation in Education and Psychology, 11*(2), 113-130. doi: 10.21031/epod.639148

Received: 28.10.2019
Accepted: 04.05.2020

education and training, workplace, home and entertainment (Mesleki Yeterlik Kurumu-MYK, 2016, p.23).

It is seen that the focus is on the skill development of individuals in student competencies defined both in the international and national fields. However, it requires a process both for a person to have knowledge or skills and to determine the knowledge and skills he or she has. Furthermore, possessing knowledge or skill does not necessarily mean that an individual can express or display that knowledge or skill. To illustrate, an individual

- may not be aware of the knowledge or skill s/he possesses,
- may not consider this knowledge essential,
- may not know how to convey this knowledge to other people (Karadüz, 2010).

Although a person may possess a skill to a certain degree, such as critical thinking, problem solving, or creative thinking, it is possible that that individual's relative standing in a group or with herself/himself over time may not be captured properly due to the limitations brought about the assessment tools used.

In traditional testing settings used for the assessment of writing proficiency, students often are asked to write a composition given a subject or the main idea, such as friendship or about some assertion, such as looking and seeing are not the same. Here it can be argued that, completing the task in this manner, students might not be able to use their creativity to the full extent and perhaps might be less motivated, and hence, willing to try out strategies, routes, etc. that may help them expand their skills that they may not even know that they had. In other words, it can be argued that measurement tools to date developed for the assessment of writing skills are often limit their focus on the end product of the students' performance rather than the process experienced leading to the end product. Moreover, it is often the case that students are commonly given a passive role through which they cannot either manage or show how they managed their writing processes (Coşkun, 2013; Oral, 2014). Ülper (2008), for example, states that it is likely that the conventional means of assessments would miss essential components of writing competence in this manner, such as creativity. Hence, it would be important for writing assessments to focus on both what the students write and how they write as well as to explore about the best conditions that can provide students an opportunity to discover their own writing skills, and to manage their own writing processes (Brown, 2001). Within the scope of the argument above, this study aimed to develop a repeated assessment tool that can be used to measure college students' advanced writing skills.

Longitudinal Tests as An Alternative Measurement Tool

In this study, a process related to the development and evaluation of an alternative measurement tool, which can be used as a *longitudinal measurement tool* because it includes repeated measurements, has been introduced. Longitudinal studies are also called development studies, and they are carried out based on data collection on the same group with repetitive measurements in order to reveal the time-dependent variables (Cohen, Manion, & Morrison, 2005; Fraenkel & Wallen, 2009). Especially in studies that examine students' achievements in class, longitudinal research needs to be highlighted (Butler & Schnellert, 2012; Richardson & Liang, 2008). Because measuring the writing skills of the students in a longitudinal way, not allowing them to measure more than one time, will provide more detailed and deeper and perhaps different results than one-time measurement. The feedback that can be provided with such information to be obtained will enable the individual to become aware of the knowledge and skills he/she has, to help him feel belonging to the process and to try various ways in the process of expression. Longitudinal measurements also provide rich information to the teacher and allow students to give effective feedback (Compton-Lilly, 2003). Therefore, longitudinal measurements in class research produce much more reliable and valid results than a single measurement (Carini, 2001; Comber & Barnett, 2003; Ekwall & Shanker, 1993; Lemke, 2005; Leslie & Caldwell, 2006). In line with all this information, it is aimed to develop a measurement tool that allows longitudinal measurements within the scope of the research.

One of the important skills that 21st century individuals should acquire during the education process is creativity, and since creativity is not acquired in a single week or time period, it shows a longitudinal feature by nature like other skills. Language skills also come to the fore in expressing high-level skills such as creativity. Wittgenstein (2005) states that the conditions in which individuals live affect the process of understanding and understanding; expresses that the same words may differ in usage and context. Because the language and expressions used by individuals are a reflection of their lifestyle, and the language used by people takes shape according to lifestyle variables. In other words, it is necessary to consider the language in its natural environment and when people say something, take into account the situations they are in and the behaviors that accompany them (Wittgenstein, 2012). Accordingly, within the scope of the research, students' language skills were carried out with a focus on writing skills in which three sub-dimensions of expression, creativity and aesthetic skills were discussed.

The fact that the skills expected from individuals in the 21st century are more complex and variable in structure compared to the previous century, brings about the change of the methods used in measuring and evaluating these skills. Because creative writing does not fit into a fenced area or a pattern; It is the process of combining emotions and thoughts with imagination and transferring to the article in a subjective way (Horng, Hong & Chanlin, 2005; Oral, 2014) and in order for this process to take place, students need tasks in which they can use their creativity and bring their feelings and thoughts together. Measurement of creativity or aesthetic success can produce biased results with tasks that are limited and clearly seen what individuals expect from them.

In this study, the story completion technique was used, but unlike the traditional story completion technique, the story was divided into four parts, similar to a mini-series consisting of four parts. Each piece is united in itself and prioritizes the psychological state of the hero. Each piece of the story can be observed by spreading the practices for four weeks and by following the adaptation of the students to the process, the way they want to be able to use their creativity and aesthetic features, in other words, as cognitive and affective skills. The ability of individuals to express their abilities about writing skills in line with their motivation, interest and desires, but nevertheless in the context of story pieces given weekly and in succession, and their changing or unchanging characteristics depending on their emotional state over time. a measuring tool that recognizes.

The Aim of the Study

The aim of the present study was to develop a longitudinal assessment tool for measuring the writing skills of college students over four consecutive weeks to be rated for their sub-skills: 1) clarity of expression, 2) creativity, and 3) aesthetics. In order to conduct reliability and validity studies, data were collected from 74 college students who volunteered to participate in the four-week administrations. A survey assessing students' weekly mood was also carried out along with these administrations. In the first week of the four, students were given the first chapter of a story and were asked to write a continuation of one to several sentences. During the following three weeks, the same direction was given for the following three chapters of the story. Each chapter in the story had integrity within itself and was about the psychological state of a protagonist. Each week, students marked how they felt that week on a scale of 1 to 5 (*very good*).

Assessment development process and student performance related research questions were:

1. What should be the components of a longitudinal writing assessment tool that can be used to measure and track college students' expression, aesthetics and creativity writing subskills?
    - How to develop the process and the tasks included?
    - How to develop the scoring rubric?

2. What are the weekly state and over the week changes in individual writing performances of the students with respect to Expression, Aesthetics and Creativity subskill dimensions? How do students' week moods affect their writing performances?

   For each subskill

   - How do students perform each week?
   - How do students' performance change over the weeks?
   - How do students' performance relate to their weekly mood?

3. How to interpret individual versus group level analysis results for the changes observed in the writing subskills (expression, aesthetics and creativity)? How do students' week moods affect their writing performance as a group?
4. What is the relationship between writing skills and weekly modes in the context of expression versus aesthetics versus creativity? Do the findings support the theoretical developments in this field?

The Significance of the Study

For the purpose of making inferences about university students' advanced writing skills, a measurement tool containing a four-week repeated assessment design was developed. In addition to a weekly mood scale, a scoring rubric was developed specifically to evaluate expression, aesthetics, and creativity subskills in the students' weekly writings. With the design, it was possible to evaluate whether and to what degree students' subskill performances were related not only to their knowledge but also to their mood, interest, and adaptation level to the writing (assessment) process. The fact that no similar study was found in the related literature adds to the significance of the present study. It is for the illustrating of an alternative measurement method to the literature in its task duties, which is named as dynamic text in the research and for the completion of stories. This is one of the other feature that strengthens the significance of this study. Moreover, the present study contributes to the literature by introducing *dynamic text* based writing assessments and scoring rubrics that were most suited for the utilized repeated assessment design.

Assumptions and Limitations

Within the scope of the study, four different parts of a single story were sent to the students who were asked to complete the story. As explained in the section on the development of the task, expert opinions were received on whether or not the different parts of the story stimulated the students to the same degree and whether the parts conveyed information in the same manner; hence, it was assumed that the parts of the story stimulated the students to the same degree.

The students produced their on-line responses every week between Sunday and Tuesday. It was assumed that the students' responses reflected their weekly standings given the constructs measured.

METHOD

During the data collection stage of the study, processes compatible with the longitudinal research design were followed. Longitudinal examinations were done to reveal the time-related variations of the variables by making repeated measures and collecting data from the same group (Cohen et al., 2005; Karasar, 2008; Fraenkel & Wallen, 2009; Büyüköztürk, Çakmak, Akgün, Karadeniz, & Demirel, 2010); examining a single group having common attributes is done to reveal the general tendency of a group and the variations and tendencies of the same individuals over a certain period of time (Fraenkel & Wallen, 2009; Büyüköztürk et al., 2010).

Participants

The current study was conducted with 74 students at Gazi University, who volunteered to take part in the four-week implementation stage during the 2016-2017 academic year. Hence, the convenience sampling, which is one of the purposeful sampling methods, was used.

Data Collection Instrument

The data of the present study were collected using an on-line platform where college students could log in every week, Sunday to Thursday, to respond to the assessment module, where they were also asked to self-rate their overall mood each week.

The implementation week of the data collection tools are presented in Table 1. The students' responses to the open-ended questions were scored independently by 5 experts. The weekly moods of the students were self-ratings and ranged from 1 to 5 (*very good*).

Table 1. Data Collection Tools and Implementation

Week of measurement	Text- (The Story of Hours and Centuries)	Text based questions	Evaluation of the week
Week 1	Section I: If only he could say "Maybe"	Text completion	Weekly mood
Week 2	Section II: The darkness of light	Text completion	Weekly mood
Week 3	Section III: Accident	Text completion	Weekly mood
Week 4	Section IV: The woman's dimple	Text completion	Weekly mood

Data Analysis

Initially, a scoring rubric was developed for the rating of the responses (continuations written by the students given the storyline of the week). Student responses were at most several sentences. With the rubric developed, the responses were coded by five separate raters initially. The Krippendorff Alpha coefficients and the Pearson correlation coefficients and Intra-Group Correlation Coefficients were calculated in order to compute inter-rater agreement and to choose the best performing raters. After these calculations were made, the three raters with the highest reliability were chosen, and student scores were computed using the ratings of these raters were used in the analyses.

Initially, the variations in students' writing skills based on the sub-dimensions of expression, aesthetics, and creativity were examined. Next, graphs were made use of, and based on expert opinions, students displaying similar score patterns were categorized into subgroups. Graphs were also constructed to identify how the weekly moods reported by the students participating in the study varied in combination with the writing skills, and the student sub-groups were labeled. Discriminant analyses were conducted to determine the validity of the students' weekly mood categorizations based on the expression, creativity, and aesthetics scores. Prior to these analyses, the assumptions of missing values, normality, multiple associations, and homogeneity of variance were examined.

After the students' individual writing skills were examined, group-based descriptive statistics were calculated. Subsequently, variance analysis in the repeated measurements was run to identify whether or not students' sub-dimension scores showed significant variance across the weeks. The data set was tested for its compatibility with the analysis in accordance with the assumptions of normality and sphericity.

RESULTS

The developmental stages and the final version of the longitudinal measurement tool developed. And piloted during the present study are presented in Figure 1. The measurement and data collection design are presented in Figure 1.

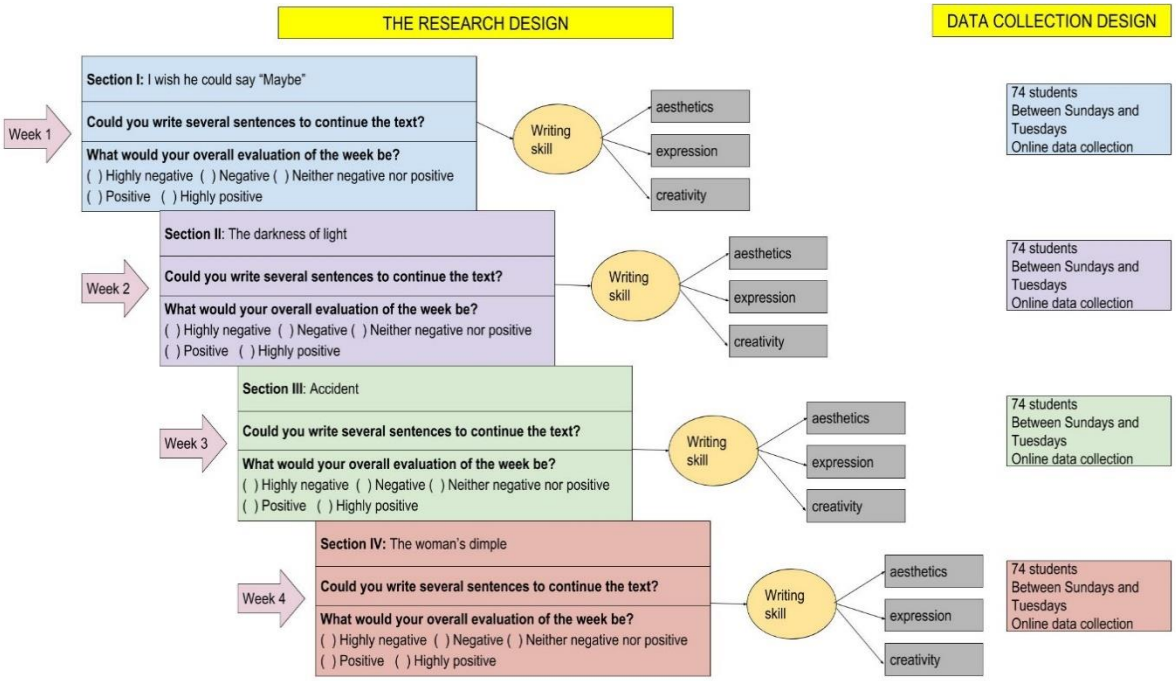


Figure 1. The Measurement and Data Collection Design

What Should Be the Components of A Longitudinal Writing Assessment Tool That Can Be Used To Measure And Track College Students' Expression, Aesthetics And Creativity Writing Subskills?

The students were sent a section of a story named as dynamic text each week and asked to complete the story texts with several sentences throughout the four weeks because meaning is a phenomenon that emerges over a course of time (Wittgenstein, 2014). The study aimed to make a longitudinal measurement of individuals' attribution of meaning to the story and what they found important within the context rather than measuring a situation in which condition of the individuals.

How to develop the process and the tasks included?

During the developmental process of the measurement tool, a *dynamic story* as a story possessing feature was created. Different from event stories, in which the introduction, development, and conclusion sections are clearly defined, situation stories are based on a central plot that is taken from one point and drawn towards a conclusion. In situation stories, the event sometimes portrays the beginning of the story, and sometimes it emerges towards the end of the story. What's important is not the event, but the effect of the protagonist upon the reader. In situation stories, also called modern stories, the understanding of time is also different as the chronological time flow is disrupted (Yakıcı, Yücel, Doğan, & Yelok, 2016). In situation stories, a pre-developed event is excluded from the story, and an analysis is made of the situation created by that event, and the events are not concluded so that the reader of the story can continue the story in his/her mind (Kolcu, 2013). Accordingly, there were two reasons why the situation story was preferred in the scope of the present study. First of all, since the research aimed to also measure such skills as creativity and aesthetics of students, which are related more to the cognitive domain, rather than having students follow up a story, the study aimed to reveal students' feelings, as well as their opinions. Secondly, as the story completion technique was used, the researchers wanted the readers to identify with the story so that they could easily complete the story.

Moreover, the story was constructed in accordance with the post-modernism movement using the technique of stream of consciousness. Accordingly, metaphors and allusions came to the fore in combining the sections forming the whole. As can be observed in Figure 1, there are seasonal transitions among the sections of the story. In addition, while the "woman's dimple" is described in

the first three sections of the story as something beautiful, at the end of the story it disappears. With the consideration of the post-modern approach, the dimple in place of the "woman's eyes" disappears as the woman's eyes open. In the story, which was titled "Hours and Centuries" as it included the night and the events experienced, there is an allusion that the hours of a night stretch out like a century. After the story was finalized and divided into four sections, three expert opinions were received from experts on New Turkish Literature on the story text (with respect to expressions, descriptions, structure, and logical sequencing). After some of the expressions were modified, opinions of the first measurement and assessment expert (as regards the items) were received. Thus, the story component of the measurement tool, named as "dynamic text", was created.

The readers found the protagonist in his room at a time in the night and watched him go through all the details in his mind of a memory that didn't allow him to sleep. Suddenly the woman in his memories approached him on a spring day and the writer was introduced to the woman in his memories. A past moment experienced within the lives of the writer, and the woman is left incomplete. In the second section of the story, the protagonist is in the later hours of the night and is then looking outside the room. Within the lights he watches outside, he suddenly finds himself within a summer day memory. The reader, who had met the woman in the previous week, gets to know another attribute of the woman this time, and the talk between the two is again left incomplete. In other words, it is inscribed into the writer's soul or his thoughts. In the third section of the story, the night is slowly leaving its place to the dawning day, but the writer is still suffering under his thoughts. This time the reader witnesses where and how the writer meets the woman on an autumn day. But everything is left incomplete. In the last section of the story, the reader sees that it is morning, a winter morning. And as the night ends and the day dawns, everything that was experienced passes through his mind, and the situation that does not make him sleep starts to appear but is not known completely. The story created within the scope of the present study is presented in Table 2.

How to develop the scoring rubric?

Each week the responses given by the students to complete the storyline were examined to see in which category they could be evaluated. Thus, initially, the related literature was reviewed to identify the sub-dimensions of higher-order writing skills, and scorings were done by five raters. A holistic scoring key was created to score sub-categories. Then, modifications were done in the sub-dimensions in the rubric based on the raters' views and the inter-rater agreement analyses. Subsequently, the sub-dimensions of the writing skill, namely, aesthetics, expression, and creativity were identified. The scoring of each sub-dimension was done within the range of 0-4, and the essential details were added. The modified version of the scoring rubric, which is scored with a holistic approach, was used by three raters (selected from the total 5, due to higher reliability) to score the four-week responses of 74 students, and the rubric was finalized. The sub-dimensions in the rubric used for scoring are as follows:

- Creativity: If the students used cliché expressions, the response received 0 point; if they used creative and unique expressions, that is developed a new perspective, their response received a score between 1 to 4 points.
- Expression: The students' responses were scored between 0 to 4 points based on suitability to the flow (suitability to the context), meaningfulness, consistency, and coherence.
- Aesthetics: Students' descriptions in their expression in the texts, elevating the experiences, and catching a worthwhile tone were scored between 0 and 4 points.

Initially, the Pearson correlation coefficient was calculated based on the raters' scores, and the coefficients for the sub-dimensions for each week were found to range between .53 and .87. It was also determined that the correlation coefficient within the group calculated according to the answers of the raters was calculated between .53 and .88. Subsequently, the calculated Krippendorff Alpha coefficients for all the sub-dimensions and the weeks were found to range between .52 and .67.

Table 2. The Stories

Week	Section
1	It's 02.25… As the room was imperceptibly dawning with the street lights, it was absorbing the entire quietness of the night. It was not even allowing him to shut his foxy, mind-craving eyes. It was not possible anyway for him to forget that day when the moist drops fell to the ground, and the weather was nice. How he had gotten excited while getting dressed. He had gotten leave from work, gotten up early and ironed his shirt, and unsatisfied, ironed it again. He had sat at a table in a tea garden, placed the flowers he brought with him on the table and then waited for quite some time. Yes, there she was, coming… There was no wind blowing in the air, no moving leaf; the universe was merely at a still. The woman with a dimple emerging on her cheek when she smiled sat across him. Gazing at her, he attempted to say, "There are things I need to explain to you…" but the woman got the jump on him; he couldn't say it. If only he could say "Maybe"…
2	It's 03.27... The room is quieter than the night... The lights of a flat from the opposite apartment building went on. Perhaps there were others, apart from himself, that could not sleep. Or they got up just to drink water, who knows? He wanted them to go to bed immediately and turn off the lights; the street lights were also to be turned off. The entire world was to be buried in the darkness like the life he was living. Why were people making such an effort to see the daylight anyway? He wished he could forget that summer daylight that caused his mind to bleed. What would happen if people did not see people rushing here and there, the fighting children while playing in the street, the men frustrated with the hot weather? They were in a hospital where there was a fan circulating on the ceiling. The woman said, "I want to see you" and again smiled with a dimple appearing on her cheek. What had the Little Prince said? "It is only in the heart that one can see rightly, what is essential is invisible to the eye. Oh, how he could explain this to her, but he couldn't. If only he could say "Maybe"…
3	It's 06.23... He became happy when the room slowly started to enlighten because it was impossible for him to justify himself during the nights when people knew his secrets. Everything slowed down during the nights; time was becoming impatient to confront the things that were experienced. He wished he could stop the time on that autumn day. The weather was warm, the leaves were yellow, people were happy... It was as if everyone were in their corner preparing for a stormy winter. If he could slow down too; why was he proceeding so fast anyway? He had just bought his car and was trying it out. It was too late when he had seen the woman crossing the street. The intermingling sounds of the brakes, his heart, and the woman suddenly came to an end. How could he forget the expression on the woman's frightened, angry, surprised, suffering eyes? The eyes of the woman instantly closed, the pieces of broken glass shed blood, and her eyes never opened again. He sincerely wanted to say "I'm sorry" but he couldn't because such an apology was not to be expressed with three words …
4	It's 08.11... Was the room cold or was his life never going to warm up again? It was morning again today. He had strolled in the room all night and watched the falling snow at times. Why hadn't the whiteness covering the entire earth cover all the misdeeds? Why would people envisage the days that they know they would never experience again? The accident he had on an autumn day, the police, prosecutor, complaints, petitions, prison, and the most important of all the woman's eyes… He found the woman after his release from the prison; the woman's eyes were closed; she had not wanted to be operated. He met her; became friends with her; he first thought he would convince her. But later, he did not have the courage to build eye-contact with her. The woman was like someone who had never lived or was impossible to live on earth. She was not beautiful; her hair was not waving into the wind; and she did not like to talk much. But when she smiled unconsciously, such a dimple appeared on her cheek that one could stay awake for days just to see it. One spring day, he wanted to explain everything. But the woman got the jump. "I forgave the person who did this to me," she said and added, "For me, there was no beauty in the world worth seeing until I met you…" One spring day, the woman had surgery, and her eyes opened. Again they came eye to eye. The woman smiled, but this time no dimple accompanying her smile appeared on her cheek…

2. What Are the Weekly State And Over the Week Changes In Individual Writing Performances Of The Students With Respect To Narration, Aesthetics And Creativity Subskill Dimensions? How Do Students' Week Moods Affect Their Writing Performances?

During the evaluation process of the measurement tool, initially individual variations (across the weeks for the same individual) and then group-based variations (each week for all the individuals in the group) in students' writing skill levels based on the repeated measurements of the sub-dimensions of expression, aesthetics, and creativity were examined.

How do students perform each week?

Table 3. The Weekly Variations in Students' Scores

Sub-Dimensions	Week	Increase		No Change		Decrease	
		f	%	*f*	%	*f*	%
Aesthetics	week 2 → week 1	42	56.7	7	9.5	25	33.8
	week 3 →week 2	27	36.5	8	10.8	39	52.7
	week 4 → week 3	35	47.3	10	13.5	29	39.2
Creativity	week 2 → week 1	42	56.8	8	10.8	24	32.4
	week 3 → week 2	29	39.2	5	6.8	40	54.1
	week 4 → week 3	38	51.4	9	12.2	27	36.5
Expression	week 2 → week 1	35	47.3	11	14.9	28	37.8
	week 3 → week 2	40	54.1	10	13.5	24	32.4
	week 4 → week 3	24	32.4	15	20.3	35	47.3

Initially, the weekly variations in the students' writing skill scores were examined (Table 3). As can be observed in Table 3, in the aesthetics sub-dimension, it was revealed that 56.7% (n = 42) of the students scored higher in the second week of the implementation when compared to the first week, while 33.8% (n = 25) received lower scores. Only 9.5% (n = 7) of the students were found to receive the same score in the first two weeks. In other words, it was revealed that 91.5% of the students' scores were found to change in the weeks following the first two weeks. The change in the scores students received in the second and third weeks was examined. It was found that there was an increase in the scores of 36.5% (n = 27) of the students, while there was a decrease in the scores of 52.7% (n = 39) of the students; the scores of 10.8% (n = 8) of the students remained the same in these two weeks. In the second and third week of the implementation, the aesthetics scores of 89.2% of the students were found to have changed. It was observed that 47.3% (n = 35) of the students participating in the study received higher scores in aesthetics in the fourth week when compared to the third week, while 39.2% (n = 29) received lower scores; 13.5% (n=10) of the students' scores remained the same. Thus, 33.8% (n = 25) of the students' aesthetics scores were found to have changed in the third and fourth weeks.

As can be observed in Table 3, in the creativity sub-dimension, it was revealed that 56.8% (n = 42) of the students participating in the study scored higher in the second week of the implementation, when compared to the first week, while 32.4% (n = 24) received lower scores. 10.8% (n = 8) of the students were found to receive the same score in the first and second weeks. In other words, it was revealed that 89.2% of the students' creativity scores were found to be different in the first and second week. It was found that 39.2% (n = 29) of the students' creativity scores were higher in the third week than they were in the second week, while 54.1% (n = 40) of the students' scores were lower. The creativity scores of 6.8% (n = 5) of the students were found to remain the same in the second and third weeks. Overall, the creativity scores of 93.2% of the students participating in the study were found to have changed throughout these weeks. The study also revealed that 51.4% (n = 38) of the students participating in the study received higher creativity scores in the fourth week when compared to the third week, while 36.5% (n = 27) received lower scores. 12.2% (n = 9) of the students were found to receive the same creativity scores in the third and fourth weeks. Overall, it can be observed that 87.8% of the students' creativity scores changed in the last two weeks.

As can be observed in Table 3, in the expression sub-dimension, it was revealed that 47.3% (n = 35) of the students scored higher in the second week of the implementation when compared to the first week, while 37.8% (n = 28) received lower scores. On the other hand, 14.9% (n = 11) of the students were found to receive the same scores in the expression sub-dimension in the first two weeks. It was also revealed that 54.1% (n = 40) of the students participating in the study received higher scores in the expression sub-dimension in the third week, when compared to the second week, while 32.4% (n = 24) received lower scores. On the other hand, 13.5% (n = 10) of the students were found to have received the same scores in the expression sub-dimension in the second and third weeks. Overall, 86.5% of the students' expression scores were observed to have changed throughout these weeks. It

was also found that 32.4% (n = 24) of the students participating in the study received higher scores in the fourth week than they did in the third week, while 47.3% (n = 35) received lower scores. The results showed that 20.3% (n = 15) of the students' expression scores remained the same in the third and fourth weeks. Overall, there was a change in the scores of 79.7% of the students participating in the study in the expression sub-dimension. In conclusion, the study revealed that the scores of students in all three sub-dimensions –Aesthetics, Expression, and Creativity– varied (increased or decreased) in the repeated measurements. As a result of the close examinations, it was observed that even though a student may have received the same total score in two measurements, the scores received in any of the sub-dimensions throughout the four weeks were not the same. In other words, there was no student receiving the same score for aesthetics, creativity, and expression throughout the four weeks. Sample responses received from the students in the scope of the present study are presented in Table 4.

Table 4. Example

Changes	Week	Text Completion
Those displaying a positive change	1	How much he loved her … the incidents could have proceeded in his favor
	2	What actually passed through his mind… everything was going to be easier for him
	3	What would have changed even if he apologized? What was done was done. Everything was going to remain the same. Perhaps my conscience was going to soothe. But it was going to continue to stay in my mind like the first day.
	4	She was smiling at me but the missing dimple was giving it away. The regret I had been feeling all this time increased more at that moment. I wished I had never been there, never met that woman. Perhaps I could not roll back the clock, but I was going to do everything I could to bring back the woman's dimple.
Those displaying a negative change	1	He shouldn't have given in to the things he heard. If only he had understood the real meaning underlying what the woman said and hadn't left her alone in the dark well, then everything would have been different. I wish pride hadn't spoken first, and I wish giving up wouldn't be so easy.
	2	If only he could say "Maybe", then he would have been saved from the gnawing feeling of blindness and lived with the hope of the bright days blinding his eyes. The issue wasn't one of seeing but, well, understanding, for instance. That is, if he had said it at work, perhaps he would have been understood. Maybe then he wouldn't have had to wait for the blindness to wipe out the dimples he yearned for from his mind and memories. Maybe then there would have been other things for both of them to wait for …
	3	Even if he said it, nothing was going to change any longer. So he preferred to remain silent. Now it was time to keep quiet, to live through the pain and regret. And that's what he did. He kept quiet.
	4	However, the man could have wanted to be reborn to see that dimple.

How do students' performance change over the weeks?

The study initially revealed that the expression, aesthetics, and creativity scores of all the students (100%; n = 74) varied throughout the weeks; none of the students' scores remained the same throughout all the weeks. Subsequently, the combined variation in the expression, aesthetics, and creativity scores was examined, which revealed that there were students scoring similar and different scores with respect to the mentioned sub-dimensions of the writing skill (Figure 2-3).

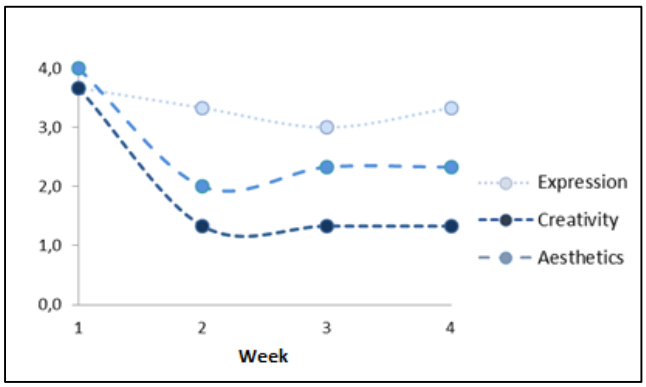


Figure 2. Different - Writing Skills Dissociated in The Context of Sub-Dimensions

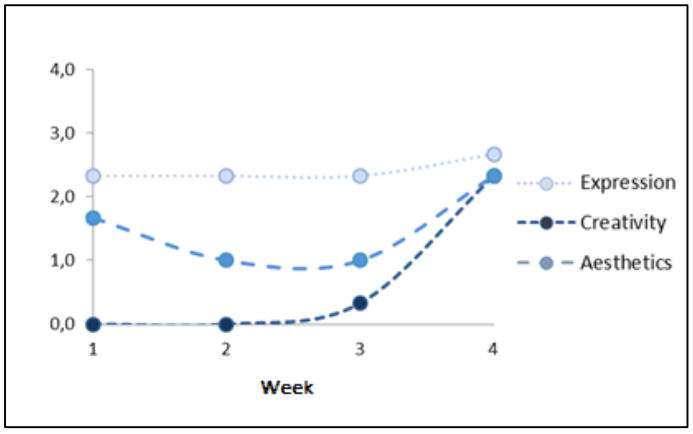


Figure 3. Similar - Writing Skills Dissociated in The Context of Sub-Dimensions

It was revealed that 17.6% (n = 13) of the students participating in the study had different scores in the sub-dimensions of the writing skill, while 31.1% (n = 23) had similar scores. It was also revealed that 51.3% (n = 38) of the students' scores varied weekly and that the variation in the sub-dimension scores showed similarity.

How do students' performance relate to their weekly mood?

Wittgenstein (2004), who stated that "Words assume meaning only within ideas and the flow of life" (p. 114), highlights that perceptions varies from one person to another, that a word or picture can be perceived and conveyed in different ways depending on the context and situation. Thus, the variation in students' writing skills was examined in combination with the change in their weekly mood. The groups that this examination yielded are presented in Figure 4-6.

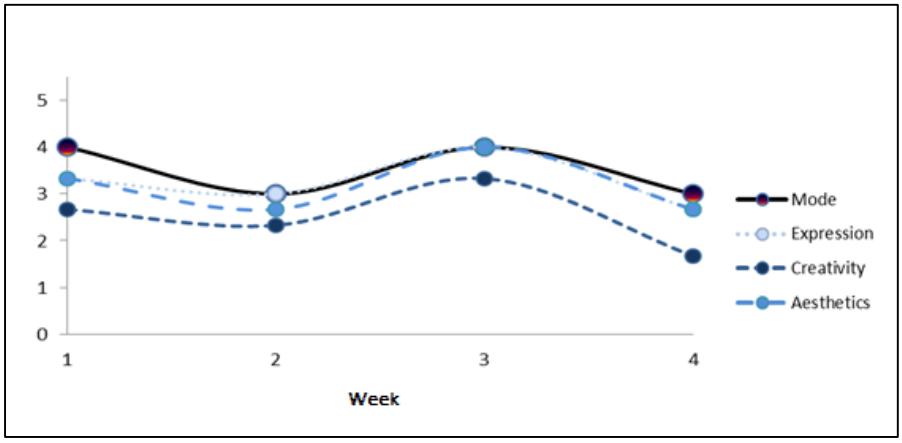


Figure 4. Similar showing an alteration of Weekly Emotional Status with Writing Skill

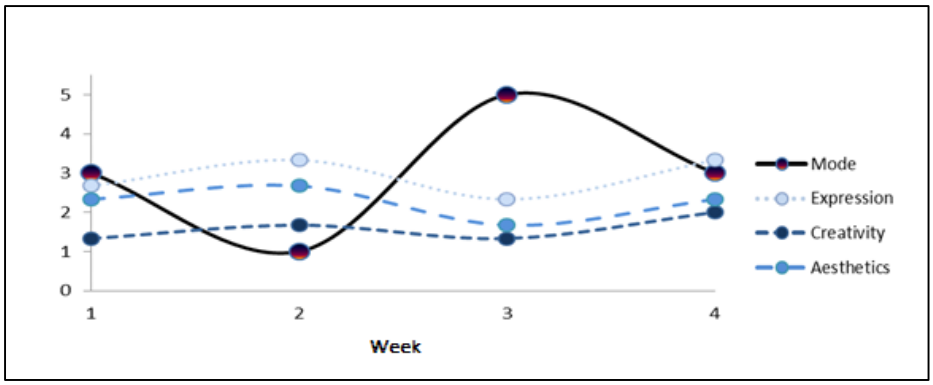


Figure 5. Opposite showing an alteration of Weekly Emotional Status with Writing Skill

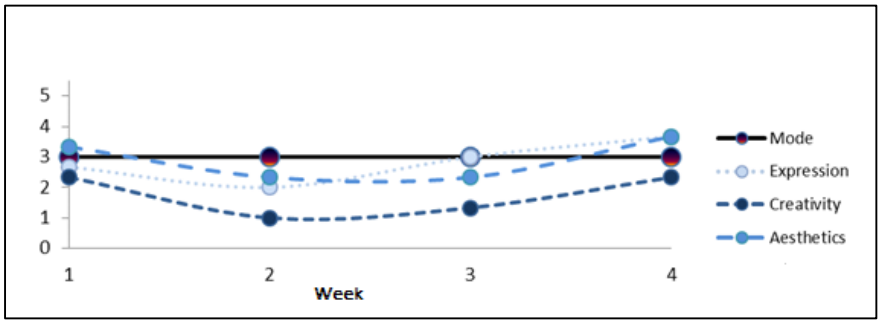


Figure 6. Different showing an alteration of Weekly Emotional Status with Writing Skill

It can be observed that the variation in the aesthetics scores of 36.5% (n = 27) of the students participating in the study show similarity in at least two weeks. The variation in the weekly mood and writing skill sub-dimensions were found to be in the negative (adverse) direction of 28.4% (n = 21) of the students in at least two weeks. The existence of a combined variation in the writing skill and weekly moods of 71.6% of the students can be considered an indication of the validity and reliability of the student responses in the measurement tool.

How to Interpret Individual Versus Group Level Analysis Results for the Changes Observed in the Writing Subskills (Expression, Aesthetics and Creativity)? How Do Students' Week Moods Affect Their Writing Performance as A Group?

The descriptive statistics of the scores the university students received from the texts based on expression, aesthetics, and creativity throughout the four weeks were calculated; the results are presented in Table 5.

Table 5. Descriptive Statistics

Dimensions	Week	N	Minimum	Maximum	$\bar{X}$	S_x	F
Aesthetics	1	74	0.00	4.00	1.93	1.03	0.96
	2	74	0.00	4.00	2.14	1.07	
	3	74	0.67	4.00	1.99	0.85	
	4	74	0.00	4.00	2.05	0.98	
Creativity	1	74	0.00	3.67	1.23	0.88	3.23*
	2	74	0.00	4.00	1.54	1.05	
	3	74	0.33	3.33	1.35	0.81	
	4	74	0.00	3.33	1.52	0.80	
Expression	1	74	0.67	4.00	2.42	0.85	4.04**
	2	74	0.00	3.67	2.54	0.86	
	3	74	1.33	4.00	2.79	0.74	
	4	74	0.00	4.00	2.59	0.84	

* $p < .05$, ** $p < .01$

Upon the examination of the information in Table 5, it can be seen that the aesthetics sub-dimension scores of the students varied between 0.00 and 4.00 during the first week of the implementation. The average of the students' aesthetics scores was calculated to be 1.93 (±1.03). The average aesthetic scores for the second, third and fourth weeks were calculated to be 2.14 (±1.07), 1.99 (±0.85), 2.05 (±0.98), respectively. Variance analysis in the repeated measures was calculated to identify whether or not there was a significant variation in the average scores received by the students in different weeks; the analysis yielded no significant variation in the students' average scores across the weeks ($F = 0.96$; $p > .05$).

It is observed that the scores the students received in the creativity sub-dimension varied between 0.00 and 3.67 in the first week of the implementation. The average creativity score of week 1 was calculated to be 1.23 (±0.88). The average creativity scores of the students for the second, third, and fourth weeks were calculated to be 1.54 (±1.05), 1.35 (±0.81), 1.52 (±0.80), respectively. The repeated measurements revealed that there was a significant variation in the students' average scores in the creativity sub-dimension across the weeks ($F = 3.23$; $p < .05$). Multiple comparison Bonferroni test was calculated in order to determine between which measurements the difference is. I. and II. Considering the errors made in multiple comparisons due to type error risks, Bonferroni test (Kayri, 2009) with the least bias was used. In order to reveal the variation, a multiple comparative Bonferroni test was utilized. As a result of this test, it was found that the scores obtained by the students in the fourth week were significantly higher than the scores they received in the first week.

With respect to the expression sub-dimension, it was revealed that the scores obtained by the students in the first week ranged between 0.67 and 4.00; their average was calculated to be 2.42 (±0.85). The students' average scores in the sub-dimension of expression for the second, third, and fourth weeks were calculated to be 2.54 (±0.86), 2.79 (±0.74), 2.59 (±0.84), respectively. As a result of the variance analysis run for the repeated measurements, there was a significant variation in the students' average scores in the sub-dimension of expression ($F = 4.041$; $p < .05$). The Bonferroni test results indicated that the average score of the students' scores in the expression dimension during the third week was significantly higher than the average score for the first week.

Upon the examination of the variation between a single group of 74 people groups, it can be observed that there was no change in the sub-dimensions of the students' writing skill or the effect of the variation was very small. In other words, it was revealed that the variations observed in students individually disappeared in the group analyses.

What Is the Relationship Between Writing Skills and Weekly Modes In The Context Of Expression Versus Aesthetics Versus Creativity? Do the Findings Support the Theoretical Developments in This Field?

In order to determine the effect of students' writing skills on weekly emotional states in repeated measurements, discrimination analysis was performed. Within the scope of the present study, initially, the frequencies and percentage values of students' weekly moods were calculated. In the first week of the implementation, the students indicated that of 2.7% (n = 2) of them had spent the week highly negatively, 17.6% (n = 13) negatively, 51.4% (n = 38) neither positively nor negatively, 25.7% (n = 19) positively, and 2.7% (n = 2) highly positively. In the second week, it was reported that 8.1% (n = 6) of the students had spent the week highly negatively, 28.4% (n = 21) negatively, 40.5% (n = 30) neither positively nor negatively, 20.3% (n = 15) positively, and 2.7% (n = 2) highly positively. In the third week, it was reported that 6.8% (n = 5) of the students had spent the week highly negatively, 5.4% (n = 4) negatively, 33.8% (n = 25) neither positively nor negatively, 41.9% (n = 31) positively, and 12.2% (n = 9) highly positively. In the last week of the implementation, the students' responses indicated that 6.8% (n = 5) of the students had spent the week negatively, 33.8% (n = 25) neither positively nor negatively, 55.4% (n = 41) positively, and 4.1% (n = 3) highly positively.

Because there was limited data, the students were categorized into two groups of negative + netural and positive weekly moods by taking into consideration the weekly reported mood scores and the average mood scores of the four weeks.

The validity of the weekly mood categorization based on students' sub-dimension scores of the writing skill was examined. The analysis was done by making calculations initially for each week and then for the four-week average scores. An average score in the context of aesthetics, expression, and creativity was obtained by first taking the average of six dimensions of students' four-week writing skills. Then, by taking the average of the emotional states they stated for each week, an average emotional score was calculated for the students. Due to the low number of data, students' emotions were classified into two groups as negative + neutral and positive. The descriptive statistics of students' writing skill scores based on weekly moods are presented in Table 6.

Table 6. Descriptive Statistics Based on Weekly Mood

Weekly Mood	Dimensions	$\bar{X}$	S_x
Negative or Neutral	Aesthetics	1.77	0.62
	Creativity	1.18	0.55
	Expression	2.42	0.55
Positive	Aesthetics	2.48	0.68
	Creativity	1.80	0.66
	Expression	2.88	0.54
Total	Aesthetics	2.03	0.73
	Creativity	1.41	0.66
	Expression	2.59	0.59

As seen in Table 6, the aesthetics, creativity, and expression skills scores of students with a good four-week average emotional state. It is observed that the mood is higher than the students with medium-low level. The Wilks' Lambda values were calculated to identify to what extent measurements based on aesthetics, creativity, and expression could distinguish students whose moods were positive and those whose moods were moderate/low; the results are presented in Table 7.

Table 7. The Wilks' Lambda Values

Sub-Dimensions	Wilks' Lambda	*F*
Aesthetics	.78	20.90***
Creativity	.79	18.65***
Expression	.85	12.37***

*** $p < .001$

When Table 7 is examined, it is seen that aesthetics, creativity, and expression skills make a significant contribution to explaining students' emotional states. Since the weekly mood variable was included in two categories in the research, a single function that clarifies 100% of the variance for the analysis of separation was calculated. The results are shown in Table 8.

Table 8. The Calculated Values Based on The Discriminant Function

Function	Eigenvalue	%Variance	%Total variance	Canonical Correlation	Wilks' Lambda	Chi-Square	df
Average	.31	100.0	100.0	.49	.76	19.03***	3

*** $p < .001$

As seen in Table 8, it is seen that the function created to separate the emotional states of the students is significant ($p < .05$). And the sub-dimensions of writing skills classify the emotional state at a

medium level (r = .49). The canonical correlation coefficients and the structure matrix correlation coefficients, calculated in relation to score types accounting for the categorization of students' four-week general average moods, are presented in Table 9.

Table 9. Standardized Canonical Correlation Coefficients and Structure Matrix Correlation Coefficients of Sub-scores

Function	Canonical	Structure Matrix
Aesthetics	.94	.97
Creativity	.42	.91
Expression	.40	.74

When the information in Table 9 is examined, it can be observed that it is the aesthetics score that accounts for the categorization of students' four-week general moods most, while it is the aesthetics score that accounts for it the least. The expected and observed values and percentages of individuals in the categorization of students' general weekly moods are presented in Table 10.

Table 10. Decisions of intersecting Categorizations Based on The Discriminant Analysis

Original number of individuals	Weekly Mood	Intersecting number of students		Total
		Negative/Neutral	Positive	
Number of students	Negative/Neutral	33	14	47
	Positive	8	19	27
%	Negative/Neutral	70.2	29.8	100.0
	Positive	29.6	70.4	100.0

As can be seen in Table 10, the writing sub-dimension scores have accurately categorized 70.2% (n = 33) of the students whose weekly moods were negative/neutral and 70.4% (n = 19) of the students whose weekly moods were positive. It was found that a total of 70.3% of the students were categorized accurately.

DISCUSSION and CONCLUSION

The primary aim of the present study was to develop an alternative measurement tool to enable the observation of students' higher-order writing skills via repeated measures and the identification of more stable or dynamic sub-dimensions and the examination of the relationships among them. Within this scope, the story completion technique was utilized to measure students' writing skills in terms of their sub-dimensions: aesthetics, expression, and creativity. Different from the story completion technique, a story was written in line with the post-modern movement, and then it was divided into four sections, ensuring that each section was unified in itself. Each section was given to the students each week so that they could complete the text by writing several sentences. Since the skill of writing tends to change by nature (Borgonovi & Pál, 2016), it must be repeatedly measured to measure the skill. Accordingly, there is a need for new measurement tools that will allow students' performances to be monitored as a growth process.

Kahraman, Akbaş and Sözer (2019) mention that longitudinal measurement models can be used for modeling systematic and controlled assessment spreads over time. İlker and Melekoğlu (2017) emphasize that longitudinal studies are needed in the study of writing skills, especially in special education. Akyüz and Doğan (2017) also mention the importance of conducting longitudinal studies in the process of an in-depth study of literacy skills. Similarly, Funder (2006) points out that the measurements of the behaviors acquired in a process should not be in one go. Since longitudinal studies allow the monitoring of the effects of the time factor (Werner, 2013; Norris, Tracy, & Galea, 2009), the measurement of the behavior gained over time also needs to be measured based on time.

Accordingly, there is a need for measurement tools that will allow students to track their writing skills over time, not at once.

In the current study, an alternative measurement tool was studied to be used to provide feedback to students with respect to their sub-writing skills. Namely, a longitudinal measurement tool was developed containing repeated assessments. The results from the conducted application study revealed that there were, in fact, variations in the writing skills of students in terms of aesthetics, expression, and creativity subskills, and this was so for most of the students over the four weeks. The changes were not always suggestive of growth. This suggests that other factors than students writing ability were playing a role in their performances. Such factors may include but are not limited to students' familiarity with the storyline, adaptation level to the task process, curiosity, motivation or well-being at the time of writing. Our results suggest that it might be that as students continued to write to the same storyline over the weeks, 1) some may have become more aware of their own potential versus productivity in writing and performed better and 2) some have lost their motivation and experienced a decline in their writing performance. Regardless of the outcome in scoring, these findings suggest that the assessment format being "longitudinal" rather than cross-sectional, it was possible to infer from the ratings that some factors other than those related to the cognitive processes alone play a role when it comes to portray advanced writing skills.

These findings exemplify the probable contributions of examining within-person variations through repeated measures while investigating between-person variations. With repeated and meaningful observations, it is particularly important to develop measurement tools in which individuals' general well-being, interest, desire, and motivation levels are taken into consideration, especially when higher-order skills are of interest. Avey, Luthans and Mhatre (2008) point out that determining the feature that is the subject of measurement based on longitudinal measurement approach shows changes and developments over time, and the observed effect sizes are important in terms of enabling emotionalization. The results obtained in this study illustrate that the weekly changes in students' moods can have a profound effect on their writing performance, not on the sub-dimension of expression but on the sub-dimension of creativity, more so for some students than others.

Even though there was a significant variation in the individual writing scores of the students participating in the study throughout the four-week period, it was observed that the between-group differences became weaker or disappeared totally over the weeks. This shows that non-linear variations can be overlooked within a group, and thus, group-based examinations can be insufficient in revealing individual-based (within-person) variations. A literature review indicates that in studies focusing on measuring the writing skill (Çıralı, 2014), analyses were often carried out with average scores over groups. Kahraman et al. (2019) report that traditional one-shot assessment tools may not be sensitive enough to capture within-person variations when cognitive or affective skills of interest are prone to change or subject to growth over time. Muthén & Curran (1997) points out that longitudinal measurement emotional states provide reliable and valid evidence for measuring and evaluating individual differences. In the application carried out within the scope of this research, it was determined that the individual changes of the students disappeared in group-based examinations.

Given the results of the present study, researchers are recommended to investigate if the construct they are interested in measuring is subject to change over time, and if so, to consider formulating a repeated assessment design, one that preferably includes relevant affective measures, such as, motivation. This way, additional valuable data may be collected to support the validity of the inferences to be made using the assessment results.

In the present study, 74 students were reached. A similar study can be conducted with more students to examine the relationship between the writing skill and its sub-dimensions and their factor load values by means of multiple group models. Within the scope of the study, an appropriate text to fit the event story was constructed. Making similar measurement tools more common via different techniques is recommended.

As it was observed in the results of the present study, individual differences might be overlooked if data analyses are based on group differences alone; so that, it is recommended that within person

differences should also be investigated, especially when there is potential for growth in students' side, that is, to track growth or support change in individuals' performance over time. This would help researchers understand the factors that may help or hinder student performance better.

During the present research, the Aesthetics, Expression, and Creativity sub-dimensions were coded in the scoring rubric. In other studies, the tone of texts, such as a sad tone, a happy ending etc. in the stories that the students complete can be examined with respect to features such as empathy and the language style of the narrator.

Acknowledgment

We thank research assistant Sebahat Gören Kaya, Derya Akbaş, Ergün Cihat Çorbacı and Esra Sözer who helped us during the data collection and scoring stages of this research.

REFERENCES

Akyüz, E., & Doğan, Ö. (2017). Ev okuryazarlık ortamı: Tanımları, boyutları ve kendiliğinden ortaya çıkan okuryazarlık becerilerinin gelişimindeki rolü. *H.Ü. Sağlık Bilimleri Fakültesi Dergisi, 4*(3), 38-57. Retrieved from http://static.dergipark.org.tr/article-download/f05d/90fc/2b9b/5a44d125671b5.pdf?

Avey, J. B., Luthans, F., & Mhatre, K. H. (2008). A call for longitudinal research in positive organizational behavior. *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior, 29*(5), 705-711. doi: 10.1002/job.517

Borgonovi, F., & Pál, J. (2016). *A framework for the analysis of student well-being in the PISA 2015 study*. Paris: OECD Publishing.

Brown, H. D. (2001). *Teaching by principles: An interactive approach to language pedagogy.* New York, NY: Addison Wesley Longman.

Butler, D. L., & Schnellert, L. (2012). Collaborative inquiry in teacher professional development. *Teaching and Teacher Education, 28*(8), 1206-1220. doi: 10.1016/j.tate.2012.07.009

Büyüköztürk, Ş., Çakmak, E. K., Akgün, Ö. E., Karadeniz, Ş., & Demirel, F. (2010). *Bilimsel araştırma yöntemleri*. Ankara: PegemA Akademi.

Carini, P. F. (2001). *Starting strong: A different look at children, schools, and standards*. New York, NY: Teachers College Press.

Çıralı, H. (2014). *Dijital hikâye anlatımının görsel bellek ve yazma becerisi üzerine etkisi* (Yayımlanmamış yüksek lisans tezi, Hacettepe Üniversitesi, Ankara). Erişim adresi: http://tez2.yok.gov.tr/

Cohen, L., Manion, L., & Morrison, K. (2005). *Research methods in education.* (5th ed.). London: Routledge Falmer.

Comber, B., & Barnett, J. (Eds.). (2003). *Look again: Longitudinal studies of children's literacy learning.* Newtown, Australia: Primary English Teaching Association.

Compton-Lilly, C. (2003). *Reading families: The literate lives of urban children*. New York, NY: Teachers College Press.

Coşkun, E. (2013). Yazma becerisi. A. Kırkkılıç ve H. Akyol (Ed.), *İlköğretimde Türkçe öğretimi* içinde (ss. 49-91). Ankara: Pegem Akademi.

Ekwall, E. E., & Shanker, J. L. (1993). *Ekwall/Shanker reading inventory*. Boston, MA: Allyn & Bacon.

European Commission. (2006). *Report promoting cultural education in Europe: A contribution to participation, innovation and quality*. Austrian Presidency.

Fraenkel, J. R., & Wallen, N. E. (2009). *How to design and evaluate research in education.* (7th ed.). New York, NY: McGraw-Hill International Edition.

Funder, D. C. (2006). Towards a resolution of the personality triad: Persons, situations, and behaviors. *Journal of Research in Personality, 40*(1), 21-34. doi: 10.1016/j.jrp.2005.08.003

Horng, J. S., Hong, J. C., & Chanlin, L. J. (2005). Creative teachers and creative teaching strategies. *International Journal of Consumer Studies, 29*(4), 352-358. doi: 10.1111/j.1470-6431.2005.00445.x

İlker, Ö., & Melekoğlu, M. A. (2017). İlköğretim döneminde özel öğrenme güçlüğü olan öğrencilerin yazma becerilerine ilişkin çalışmaların incelenmesi. *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Özel Eğitim Dergisi, 18*(3), 443-469. doi: 10.21565/ozelegitimdergisi.318602

Kahraman, N., Akbaş, D., & Sözer, E. (2019). Bilişsel olmayan öğrenme durum ve süreçlerini ölçme ve değerlendirmede boylamsal yaklaşımlar: Duygu cetveli uygulamasi örneği. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi, 19*(1), 257-269. doi: 10.17240/aibuefd.2019.19.43815-459831

Karadüz, A. (2010). Dil becerileri ve eleştirel düşünme. *Turkish Studies 5*(3), 1566-1593. doi: 10.7827/TurkishStudies.1572

Karasar, N. (2008). *Bilimsel araştırma yöntemleri*. İstanbul: Nobel Yayın Dağıtım.

Kayri, M. (2009). Araştırmalarda gruplar arası farkın belirlenmesine yönelik çoklu karşılaştırma (post-hoc) teknikleri. *Fırat Üniversitesi Sosyal Bilimler Dergisi, 19*(1), 51-64, http://web.firat.edu.tr/sosyalbil/dergi/arsiv/cilt19/sayi1/051-064.pdf adresinden edinilmiştir.

Kolcu, A. İ. (2013). *Öykü sanatı*. Erzurum: Salkımsöğüt Yayınları.

Lemke, J. (2005). Place, pace, and meaning: Multimedia chronotopes. In S. Norris & R. H. Jones (Eds.), *Discourse in action: Introducing mediated discourse analysis* (pp. 110-122). New York, NY: Routledge

Leslie, L., & Caldwell, J. (2006). *Qualitative reading inventory*. Reading, MA: Allyn & Bacon.

Muthén, B. O., & Curran, P. J. (1997). General longitudinal modeling of individual differences in experimental designs: A latent variable framework for analysis and power estimation. *Psychological Methods, 2*(4), 371-402. Retrieved from https://www.jhsph.edu/research/centers-and-institutes/johns-hopkins-center-for-prevention-and-early-intervention/Publications/muthen.curran.1997.pdf

Mesleki Yeterlik Kurumundan; Türkiye Yeterlilikler Çerçevesine Dair Tebliğ. (Tebliğ No: 2015/1). Türkiye Yeterlilikler Çerçevesi. *Resmî Gazete*, 29581, 02.01.2016

Norris, F. H., Tracy, M., & Galea, S. (2009). Looking for resilience: Understanding the longitudinal trajectories of responses to stress. *Social Science & Medicine*, *68*(12), 2190-2198. doi: 10.1016/j.socscimed.2009.03.043

Oral, G. (2014). *Yine yazı yazıyoruz*. Ankara: Pegem Akademi Yayıncılık.

Richardson, G. M., & Liang, L. L. (2008). The use of inquiry in the development of preservice teacher efficacy in mathematics and science. *Journal of Elementary Science Education, 20*(1), 1-16. doi: 10.1007/BF03174699

Ülper, H. (2008). *Bilişsel süreç modeline göre hazırlanan yazma öğretimi programının öğrenci başarısına etkisi* (Doktora tezi, Ankara Üniversitesi Sosyal Bilimler Enstitüsü, Ankara). Erişim adresi: http://tez2.yok.gov.tr/

Werner, E. E. (2013). What can we learn about resilience from large-scale longitudinal studies? In S. Goldstein & R. B. Brooks (Eds.), *Handbook of resilience in children* (pp. 91-105). Boston, MA: Springer US

Wittgenstein L. (2004). *Tractatus logico-philosophicus*. London: Routledge & Kegan Paul.

Wittgenstein, L. (2005). *Philosophical grammar*. USA: University of California Press.

Wittgenstein, L. (2012). *Philosophical investigations*. UK: Blackwell Publishing.

Wittgenstein, L. (2014). *Recherches philosophiques*. Paris: Editions Gallimard.

Yakıcı, A., Yücel, M., Doğan, M., & Yelok, S. (2016). *Yazılı anlatım üniversiteler için Türkçe-1*. Ankara: Yargı Yayınevi.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2020; 11(2); 131-146

An Evaluation of 4PL IRT and DINA Models for Estimating Pseudo-Guessing and Slipping Parameters *

Ömür Kaya KALKAN ** İsmail ÇUHADAR ***

Abstract
In an achievement test, the examinees with the required knowledge and skill on a test item are expected to answer the item correctly while the examinees with a lack of necessary information on the item are expected to give an incorrect answer. However, an examinee can give a correct answer to the multiple-choice test items through guessing or sometimes give an incorrect response to an easy item due to anxiety or carelessness. Either case may cause a bias estimation of examinee abilities and item parameters. Four-parameter logistic item response theory (4PL IRT) model and the deterministic inputs, noisy, and gate (DINA) model can be used to mitigate these negative impacts on the parameter estimations. The current simulation study aims to compare the estimated pseudo-guessing and slipping parameters from the 4PL IRT model and the DINA model under several study conditions. The DINA model was used to simulate the datasets in the study. The study results showed that the bias of the estimated slipping and guessing parameters from both 4PL IRT and DINA models were reasonably small in general although the estimated slipping and guessing parameters were more biased when datasets were analyzed through the 4PL IRT model rather than the DINA model (i.e., the average bias for both guessing and slipping parameters = .00 from DINA model, but .08 from 4PL IRT model). Accordingly, both 4PL IRT and DINA models can be considered for analyzing the datasets contaminated with guessing and slipping effects.

Key Words: 4PL IRT model, DINA model, (pseudo) guessing effect, slipping effect, lower-upper asymptote parameter.

INTRODUCTION

Psychological and educational tests are usually used for observing a sample of examinees' behaviors. Many of them focus on measuring the abilities and skills of examinees. Therefore, it is important to know how an examinee's ability determines the correctness of an answer on an item (Lord, 2012). In an achievement test, a correct response is expected from an examinee with the required knowledge on the item whereas an examinee without the necessary knowledge on the item is supposed to give an incorrect answer (Rowley & Traub, 1977). However, this assumption may not hold for the multiple-choice test items. In a test with multiple-choice test items, an examinee's response may be a reflection of true ability, guessing behavior or unexpected incorrect response (i.e., slipping effect) due to anxiety or carelessness (Liao, Ho, Yen, & Cheng, 2012; Yen, Ho, Laio, Chen, & Kuo, 2012). Under the presence of guessing and slipping effects, the estimation of examinees' abilities and item parameters might be biased. These two effects can be modeled using item response theory (IRT) models and cognitive diagnostic models (CDMs). IRT models explain the relationship between an examinee's observed test performance and its underlying latent abilities through a mathematical function (Hambleton & Swaminathan, 1985). On the other hand, CDMs are used for determining whether an examinee has a set of attributes in order to solve a problem correctly in a test (de la Torre, 2009). CDMs have many common aspects with IRT models. For example, Junker (2001), used deterministic inputs, noisy, and gate (DINA; Haertel, 1989; Junker & Sijtsma, 2001) models as an initial tool for

* We declare that a part of this study was presented as an oral presentation at the 6th International Congress on Measurement and Evaluation in Education and Psychology (CMEEP 2018) held on 5-8 September 2018 in Prizren, Kosovo.

** Assist. Prof., Pamukkale University, Faculty of Education, Denizli-Turkey, kayakalkan@pau.edu.tr, ORCID ID: 0000-0001-7088-4268
*** Ph.D., Ministry of National Education, Ankara-Turkey, ismail.cuhadar@gmail.com, ORCID ID: 0000-0002-5262-5892

To cite this article:
Kalkan, Ö. K., & Çuhadar, İ. (2020). An evaluation of 4PL IRT and DINA models for estimating pseudo-guessing and slipping parameters. *Journal of Measurement and Evaluation in Education and Psychology, 11*(2), 131-146. doi: 10.21031/epod.660273

Received: 16.12.2019
Accepted: 02.04.2020

proposing a nonparametric IRT (NIRT) for CDMs. In addition, Junker and Sijtsma (2001) showed that, as a CDM, DINA and noisy, inputs, deterministic and gate (NIDA; Maris, 1999; Junker & Sijtsma, 2001) models meet the standard assumptions of generalized multidimensional IRT models. Similarly, Meng, Xu, Zhang, and Tao (2019) showed that four-parameter logistic (4PL) (Barton & Lord, 1981) model is a special case of the higher-order DINA model with an only one latent attribute. In addition, the authors indicated that the upper asymptote in 4PL model (i.e., d_j) corresponds to the slipping parameter in CDMs (i.e., 1 - d_j). Furthermore, Culpepper (2016) stated that the lower asymptote (i.e., *c* parameter) and the upper asymptote (i.e., *d* parameter) in 4PL IRT model correspond to the guessing and slipping parameters in CDMs, respectively. Accordingly, 4PL and DINA models including (pseudo) guessing-guess and inattention-slip parameters are described shortly in the next section.

The DINA Model

DINA model, proposed by Junker and Sijtsma (2001), requires configuring a Q matrix (Tatsuoka, 1983) as the other CDM models do. This matrix is composed of ($J \times K$ times) 1 and 0s, including attributes in the columns and items in the rows of the matrix. The element in the jth row and kth column of the matrix is showed as q_{jk}. If q_{jk} equals 1, it means an examinee is required to possess the corresponding attribute in order to answer the item correctly. If the attribute is not required for answering the item correctly, q_{jk} becomes 0 in the Q matrix. Assume vector y_i represents the observed score of an examinee *i* to *J* items and the elements of y_i are statistically independent of the required attributes vector for the test $\alpha_i = \{\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{iK}\}$. Using Q-matrix and respondent's skills vector, DINA model produces the η_{ij} in Equation 1.

$$\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}} \tag{1}$$

In Equation 1, if an examinee possesses all necessary attributes for the correct answer on the item, $\eta_{ij} = 1$; otherwise, $\eta_{ij} = 0$. DINA model allows an examinee possessing all required attributes to miss an item (slip) or an examinee without at least one of the required attributes to answer the item correctly (guess). DINA model includes a guess (*g*) and slip (*s*) parameter for each test item. The parameter g_j is defined by $g_j = P(Y_{ij} = 1 \mid \eta_{ij} = 0)$, and the parameter s_j is defined by $s_j = P(Y_{ij} = 0 \mid \eta_{ij} = 1)$. Accordingly, the probability of correct response on item j given an examinee i with an attribute profile α_i is formulated as in Equation 2.

$$P(Y_{ij}=1|\alpha) = (1-s_j)^{\eta_{ij}} g_j^{1-\eta_{ij}} \tag{2}$$

DINA model can be implemented in computer software programs, including OxEdit (Doornik, 2018), LatentGold (Vermunt & Magidson, 2016), Mplus (Muthén & Muthén, 1998-2017), "CDM" package (Robitzsch, Kiefer, George, & Uenlue, 2019) and "GDINA" package (Ma & de la Torre, 2020) available as R program (R Core Team, 2017). However, it is essential to emphasize that the implementation of the DINA model is not limited to these computer software programs.

The 4PL IRT Model

Barton and Lord (1981) proposed 4PL IRT to model a parameter for the upper asymptote in the item characteristic curve. This model accounts for unexpected incorrect responses (missing) of examinees with a high ability level due to anxiety and carelessness. In the general form of this model, the probability of correct response given the ability level is formulated as in Equation 3.

$$P[X_{ij} = 1 \mid \Theta = (\theta_1, \ldots, \theta_k), a_j, b_j, c_j, d_j] = c_j + (d_j - c_j) \frac{e^{(a_{j1}\theta_1 + \ldots + a_{jk}\theta_k) - b_j}}{1 + e^{(a_{j1}\theta_1 + \ldots + a_{jk}\theta_k) - b_j}} \tag{3}$$

In Equation 3, X_{ij} is the observed score of an examinee *i* on item *j*, *k* is the number of latent factors, Θ is the vector of examinee abilities, c_j is the pseudo-guessing parameter of item *j*, d_j is the upper asymptote parameter (i.e., slipping parameter) of item *j*, a_{jk} is the discrimination parameter of item *j*

on the latent factor k, and b_j is the intercept of item j, which is the multiplication of item discrimination and item difficulty (see Barton & Lord, 1981; de Ayala, 2009). Although Barton and Lord (1981) proposed using a common upper asymptote across all test items, the general form of the 4PL model allows estimating a different upper asymptote for each test item. One-, Two-, and Three-Parameter Logistic (1PL, 2PL, and 3PL) IRT models for dichotomous items have attracted great attention in the last decade (Magis, 2013). On the other hand, 4PL IRT model was not a commonly used IRT model among practitioners and researchers until recent years due to no indication for the benefit of using 4PL IRT model, the difficulties with the estimation of upper asymptote, and the unavailability of computer software programs that can be accessed by practitioners and researchers for using 4PL IRT model (Barton & Lord, 1981; Hambleton & Swaminathan, 1985; Loken & Rulison, 2010). However, the 4PL IRT model has become more popular in recent years, especially in the literature on IRT and computerized adaptive testing (CAT), with the development of very powerful computer software programs such as the "mirt" package in R program (Chalmers, 2012; Magis, 2013; Meng et al., 2019). Many studies have contributed to the improvement of the 4PL IRT model regarding its application in the field and parameter estimation (e.g., Culpepper, 2016; Liao et al., 2012; Loken & Rulison, 2010; Magis, 2013; Meng et al., 2019; Rulison & Loken, 2009; Yen et al., 2012).

Although the conventional IRT models allow test-takers' abilities to be scaled and ordered in one or more continuous latent factors, these IRT models including 4PL IRT model are not useful to assess test-takers' strengths and weaknesses in the latent factors because IRT models do not tell if some behaviors related to the latent factors (attributes) are mastered. Unlike IRT models, CDMs were basically proposed with the purpose of identifying test-takers' strengths and weaknesses through assessing the presence or absence of several necessary attributes to solve the problems in a test (de la Torre, Hong, & Deng, 2010; de la Torre & Lee, 2010). Among CDMs, the DINA model (Junker & Sijtsma, 2001) is a commonly used model in practice and research (DeCarlo, 2011; de la Torre, 2008). Its simple and easily interpretable formula provides a good model-data fit (de la Torre & Douglas, 2008; de la Torre & Lee, 2010). Both the 4PL IRT model and the DINA model allow *c-g* and *d-s* parameters for modeling the guessing and slipping effects, respectively.

Although the literature has many studies investigating the important factors for the estimation of item parameters accurately in IRT models and CDMs separately, there are only a few studies directly comparing the item parameters from IRT models and CDMs in the same research (e.g., 2PL vs. pG-DINA in Yakar, 2017). In addition, there are some studies employing the 4PL IRT model within the CAT (e.g., Liao et al., 2012; Yen et al., 2012). However, it is also important to investigate the parameter recovery in the 4PL IRT model for a fixed (non-adaptive) test via a simulation study because the fixed tests are commonly used in educational and psychological assessments. When the similarity between IRT models and DINA model, a restricted latent model, is taken into consideration (Culpepper, 2016; Hoijtink & Molenaar, 1997; Junker, 2001; Junker & Sijtsma, 2001; Meng et al., 2019), the current study may be helpful for the field to show the similarities and differences between 4PL IRT model and DINA model, and the important study design factors for the accurate estimation of the guessing and slipping parameters. Accordingly, the current simulation study aims to compare the estimated *c-g* and *d-s* parameters from the 4PL IRT model and the DINA model using the simulated datasets through the DINA model under several study conditions.

METHOD

Simulation Study Design

All data were generated and analyzed in the R program (R Core Team, 2017). DINA model was used for data generation. In the literature, the test length was usually between 20 and 40 in many studies (e.g., Chiu, 2008; de la Torre, 2008, 2009, 2011; de la Torre & Douglas, 2004, 2008; de la Torre & Lee, 2010, 2013; Henson & Douglas, 2005). In the data generation, test length was fixed as $J = 20$ or 40 items considering these studies in the literature. The review of the literature also showed that the

studied *g* and *s* parameters tend to be between .0 and .45 (e.g., Chiu, 2008; de la Torre & Douglas, 2004; de la Torre et al., 2010; DeMars, 2007; Henson & Douglas, 2005; Huebner & Wang, 2011). In addition, the intervals of these parameters corresponding to the low, moderate, and high levels were different across the studies. In this study, three levels of *g* and *s* parameters were manipulated in the data generation: .0 - .15 (low), .15 - .30 (moderate), and .30 - .45 (high). Then, these levels were crossed between *g* and *s* parameters in the data generation. The values of *g* and *s* parameters were equally spaced with an increment of .0075 and .00375 for the conditions with 20 and 40 items, respectively. Specifically, these values were obtained taking the ratio of intervals to test length (e.g., for the test with 20 items and the parameter values between .0 and .15, .15/20 = .0075). Then, the values of *g* and *s* parameters were fixed to $g = s = .0075$ for the first item, .015 for the second item, and .15 for the last item when test length was 20, and both *g* and *s* parameters were low (.0 - .15) in the data generation. Different values were chosen for the level of correlation among factors/attributes corresponding to the weak, moderate, and strong correlations across different studies in the literature. In this study, the correlation among the attributes was fixed to $r = .2$ (weak), .5 (moderate) or .8 (strong) considering the studies by Finch (2010), and Finch, Habing, and Huynh (2003). The chosen sample size was 500, 1000, or 2000 in some simulation studies in the literature (e.g., de la Torre 2009; de la Torre & Douglas, 2004, 2008; de la Torre & Lee, 2010, 2013). However, a sample size of 1000 is sufficiently large to use the DINA model (de la Torre et al., 2010). For the 4PL IRT model, Meng et al. (2019) used a sample size of 2000. In addition, Waller and Feuerstahler (2017) found that a minimum sample size of 1000 is necessary to obtain accurate ability estimates in the 4PL IRT model. Therefore, in this study, the sample size was fixed to $N = 3000$ considering the adequacy of the sample size for the convergence of parameters to a solution. The number of attributes is usually between 4 and 8 in the literature (e.g., Chiu, 2008; de la Torre, 2011; de la Torre & Douglas, 2004; de la Torre & Lee, 2010; Huebner & Wang, 2011). Because there were many simulation conditions included in this study and the use of a great number of attributes in a simulation study is very time consuming (de la Torre & Douglas, 2004), the number of attributes was fixed to $K = 3$ or 5. Four different Q-matrices were used in the data generation (2 test lengths x 2 different numbers of attributes). Each item was linked to one attribute in all Q-matrices (one-attribute items), and the number of items was distributed across the attributes as evenly as possible. Overall, there were a total of 108 conditions for data generation (3 *g* levels x 3 *s* levels x 3 correlation levels x 2 test lengths x 2 numbers of attributes). The number of replications for each condition was 100.

Data Analysis

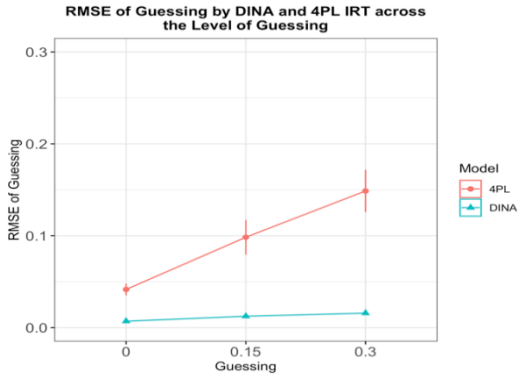
Each dataset was analyzed using a multidimensional 4PL IRT model and a DINA model. Before the analysis of datasets using the multidimensional 4PL IRT model, the dimensionality of datasets was investigated via Factor 9.2 (Lorenzo-Seva & Ferrando, 2006). Parallel analysis with the tetrachoric correlation indicated that the dimensionality assumption was met for the use of the multidimensional IRT model (i.e., it was in line with the factor structure of the datasets in the data generation via DINA model). The local independence assumption was assumed to be met because it is not within the scope of this study. Expectation-maximization (EM) algorithm was used to estimate the item parameters through 4PL IRT and DINA models because it was the default estimation method in the R packages that were used for 4PL IRT and DINA models in the study. Specifically, the analysis of datasets was conducted in the "CDM" package (Robitzsch et al., 2019) for the DINA model and the "mirt" package (Chalmers, 2012) for the 4PL model available in R program. Item-parameter bias and root mean square error (RMSE) were used to evaluate 4PL IRT and DINA models in terms of the estimation of *c-g* and *d-s* parameters correctly. 4PL IRT model was assumed to have the same true slipping and guessing parameters with the DINA model in the calculation of bias and RMSE considering the relationship between the 4PL IRT model and CDMs (see Culpepper, 2016; Meng et al., 2019). The average bias and RMSE were reported with their 95% confidence intervals across the study conditions using the formula in Equation 4.

$$\bar{\varepsilon} \pm 1.96 \frac{S_\varepsilon}{\sqrt{r_\varepsilon}} \quad (4)$$

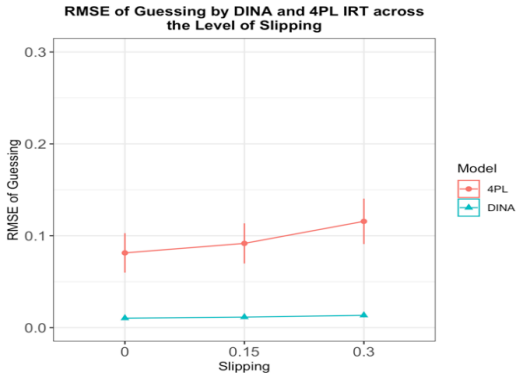
In Equation 4, $\bar{\varepsilon}$ is the average bias/RMSE of the item parameters, S_{ε} is the standard deviation of bias/RMSE of the item parameters, and r_{ε} is the number of study conditions when calculating the average bias/RMSE of the item parameters.

RESULTS

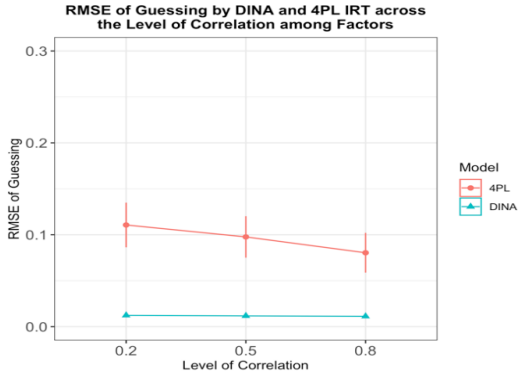
Results were summarized using the average RMSE of the item parameters and creating its 95% confidence intervals by the 4PL IRT and DINA models across the study conditions. The RMSE of guessing parameters are presented across 4PL and DINA models in Figure 1. The RMSE of the guessing parameters were almost zero across all levels of *c*-*g* parameters (*c*-*g* parameters = .0, .15, and .3; see Figure 1a), all levels of *d*-*s* parameters (*d*-*s* parameters = .0, .15, and .3; see Figure 1b), all levels of the correlation among factors/attributes (r = .2, .5, and .8; see Figure 1c), all numbers of attributes (K = 3 and 5; see Figure 1d), and all test lengths (J = 20 and 40; see Figure 1e) in the study when DINA model was fit to the data.



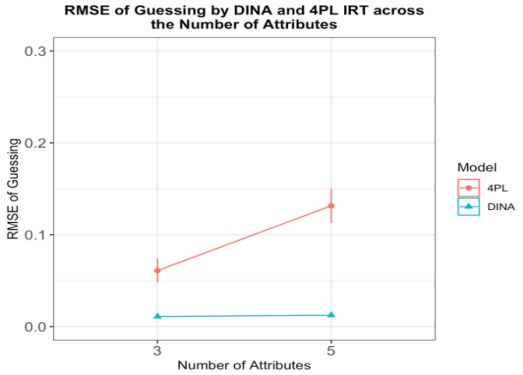
(a)



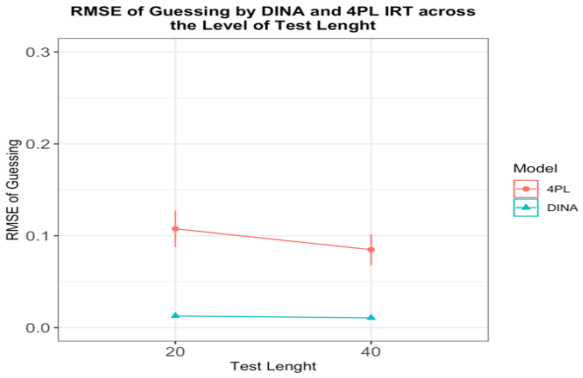
(b)



(c)



(d)

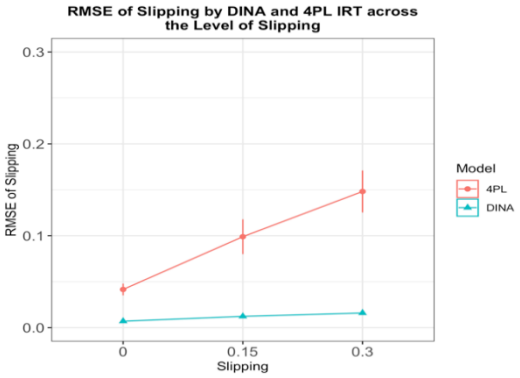


(e)

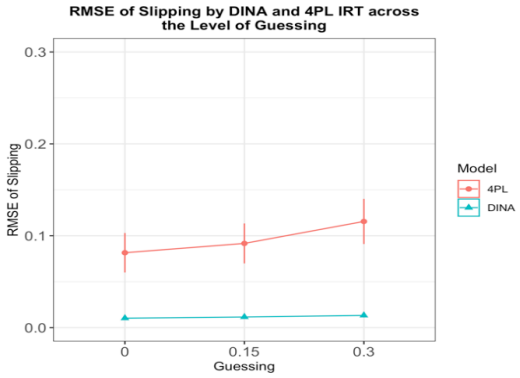
Note. On x axis of Figure 1a and 1b, 0 = the values between .0 and .15 (low); 0.15 = the values between .15 and .30 (moderate); 0.3 = the values between .30 and .45 (high).

Figure 1. The 95% Confidence Intervals of (Pseudo) Guessing-parameter RMSE by DINA and 4PL IRT Models across Different Study Conditions

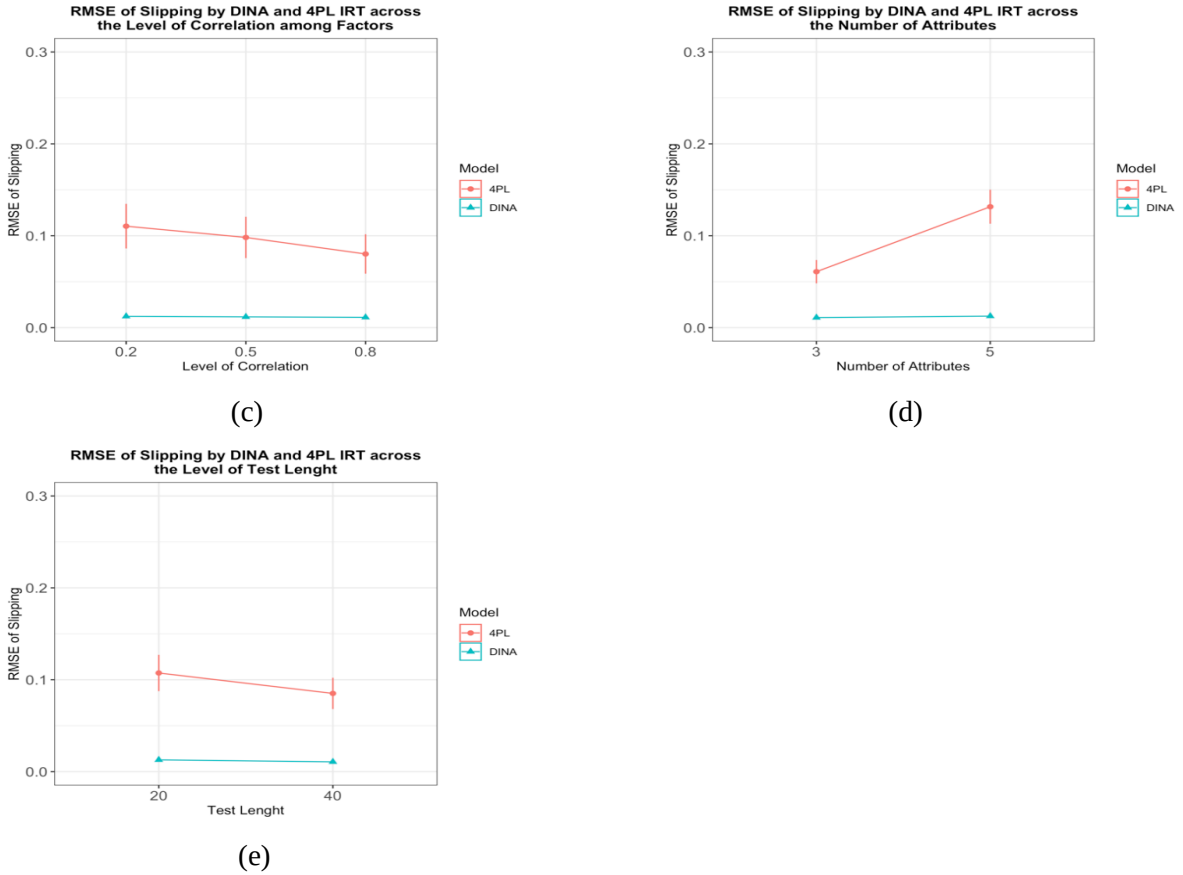
In addition, its 95% confidence intervals were so small across all these study conditions that they did not appear in any figure for DINA models. However, the average RMSE of the guessing parameters became larger across all study conditions when the 4PL IRT model was fit to the data in lieu of the DINA model (see Figure 1a, 1b, 1c, 1d, and 1e). Furthermore, the RMSE of the guessing parameters were larger for 4PL IRT model under the conditions with a larger *c-g* parameter in the data generation (the 95% confidence interval of the average RMSE for the guessing parameters was between .04 and .05 when *c-g* parameters = .0, between .08 and .12 when *c-g* parameters = .15, and between .13 and .17 when *c-g* parameters = .3; see Figure 1a). Similarly, for 4PL IRT model, the average RMSE of the guessing parameters became larger when the number of factors/attributes was greater, the test was shorter, *d-s* parameters were higher, and the correlation among factors/attributes was weaker, as expected (see Figure 1b, 1c, 1d, and 1e). However, among these four study conditions, the number of factors/attributes was the only significant study condition for the size of the RMSE of the guessing parameters from 4PL IRT model when the overlap between the 95% confidence intervals was considered (the 95% confidence interval of the average RMSE for the guessing parameters was between .05 and .07 when K = 3, and between .11 and .15 when K = 5; see Figure 1d). Overall, the similar results were also found for the RMSE of the slipping parameters (see Figure 2).



(a)



(b)

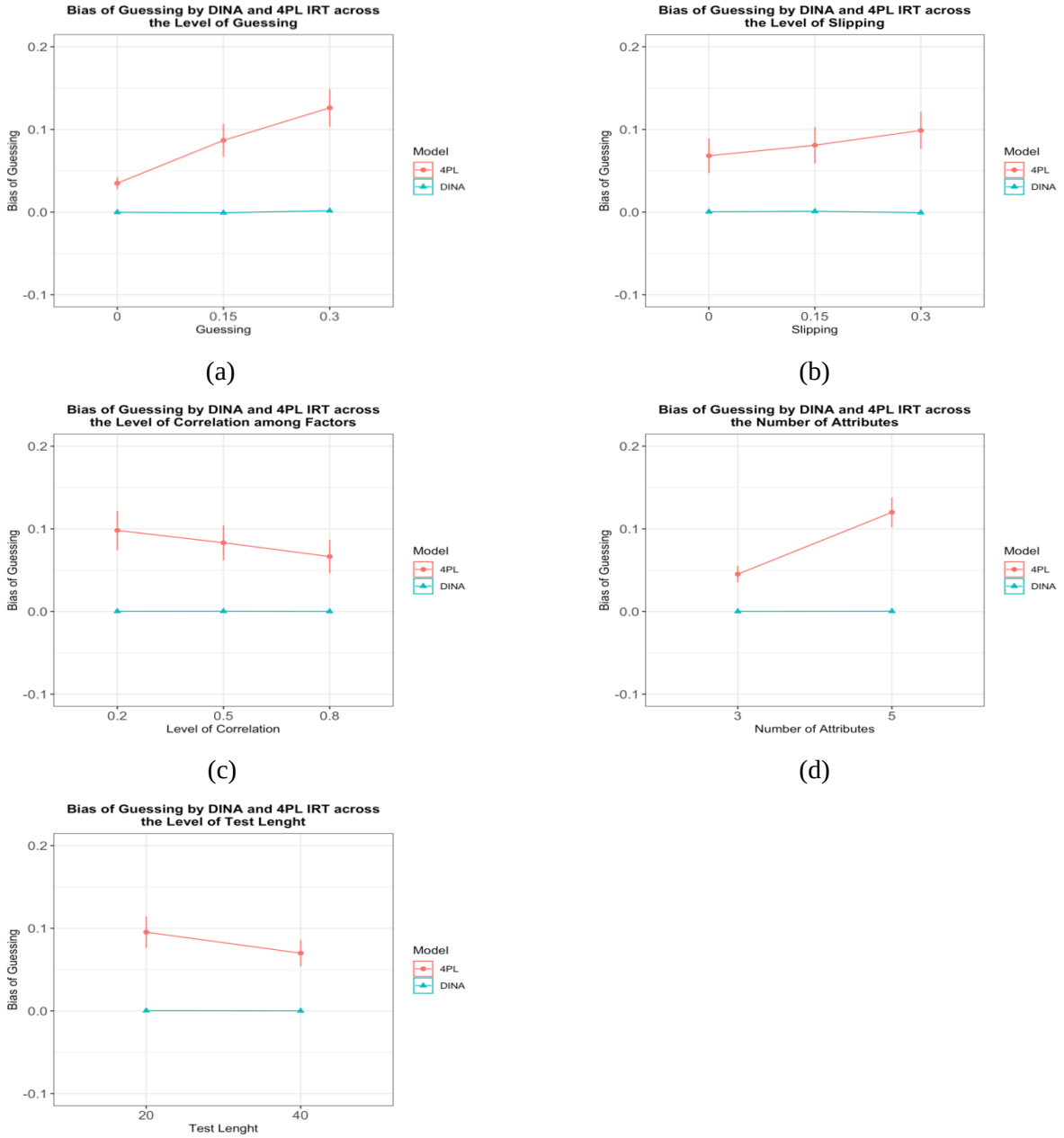


Note. On x axis of Figure 2a and 2b, 0 = the values between .0 and .15 (low); 0 .15 = the values between .15 and .30 (moderate); 0.3 = the values between .30 and .45 (high).

Figure 2. The 95% Confidence Intervals of Slipping-parameter RMSE by DINA and 4PL IRT Models across Different Study Conditions

The average RMSE of the slipping parameters with its confidence interval was almost identical to the RMSE of the guessing parameters across all study conditions for both DINA and 4PL IRT models with one exception (see Figure 2b, 2c, 2d, and 2e). The RMSE of the slipping parameters became larger for 4PL IRT model under the conditions with a larger *d-s* parameter rather than *c-g* parameter in the data generation considering the confidence intervals of average RMSEs across the study conditions (the 95% confidence interval of the average RMSE for the slipping parameters was between .04 and .05 when *d-s* parameters = .0, between .08 and .12 when *d-s* parameters = .15, and between .13 and .17 when *d-s* parameters = .3; see Figure 2a).

The bias of the guessing and slipping parameters were calculated as the expectation of the difference between the item parameters estimated from DINA or 4PL IRT models and their corresponding values from the true model in the data generation. Results were summarized using the average bias of the item parameters and creating its 95% confidence intervals by 4PL IRT and DINA models across the study conditions. The bias of guessing parameters are presented across 4PL and DINA models in Figure 3.

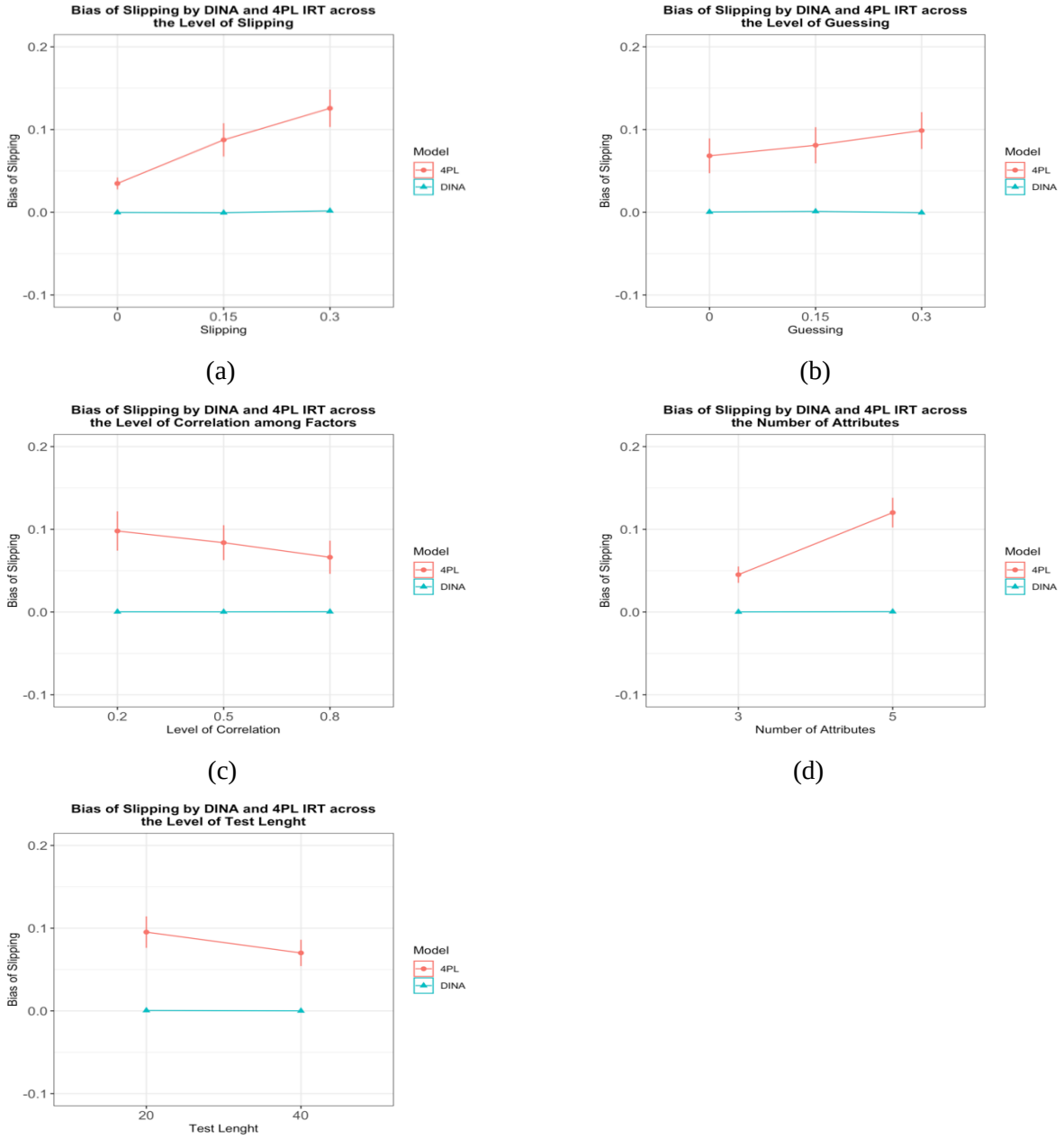


Note. On x axis of Figure 3a and 3b, 0 = the values between .0 and .15 (low); 0.15 = the values between .15 and .30 (moderate); 0.3 = the values between .30 and .45 (high).

Figure 3. 95% Confidence Intervals of (Pseudo) Guessing-parameter Bias by DINA and 4PL IRT Models across Different Study Conditions

As expected from the RMSEs of the guessing parameters, when the guessing parameters were estimated through DINA model, the bias of the guessing parameters were almost zero with a very narrow confidence interval across all levels of *c-g* parameters (*c-g* = .0, .15, and .3; see Figure 3a), all levels of *d-s* parameters (*d-s* = .0, .15, and .3; see Figure 3b), all levels of the correlation among factors/attributes (*r* = .2, .5, and .8; see Figure 3c), all numbers of attributes (*K* = 3 and 5; see Figure 3d), and all test lengths (*J* = 20 and 40; see Figure 3e) in the study. Unlike the DINA model, the guessing parameters were overestimated across all study conditions when the 4PL IRT model was used to estimate the guessing parameters (see Figure 3a, 3b, 3c, 3d, and 3e). In addition, the overestimation of the guessing parameters became more severe for the 4PL IRT model under the

conditions with a higher *c-g* parameter, a higher *d-s* parameter, a weaker correlation among factors/attributes, a greater number of factors/attributes, and a shorter test in the data generation.



Note. On x axis of Figure 4a and 4b, 0 = the values between .0 and .15 (low); 0.15 = the values between .15 and .30 (moderate); 0.3 = the values between .30 and .45 (high).

Figure 4. 95% Confidence Intervals of Slipping-parameter Bias by DINA and 4PL IRT Models across Different Study Conditions

However, among these study conditions, the value of *c-g* parameter and the number of factors/attributes in the data generation were the only study conditions that made a significant difference on the bias of the guessing parameters from 4PL IRT model considering the overlap between the 95% confidence intervals (the 95% confidence interval of the average bias for guessing

parameters was between .03 and .04 when *c-g* parameters = .0, between .07 and .11 when *c-g* parameters = .15, and between .10 and .15 when *c-g* parameters = .3; between .04 and .05 when K = 3, and between .10 and .14 when K = 5; see Figure 3a and Figure 3d, respectively). The similar results were also found for the bias of the slipping parameters (see Figure 4). However, like the RMSE of the slipping parameters, the overestimation of the slipping parameters were more severe under the conditions with a larger *d-s* parameter rather than a larger *c-g* parameter in the data generation when the 95% confidence intervals of the average bias for the slipping parameters were taken into consideration across the study conditions (i.e., the 95% confidence interval of the average bias for slipping parameters was between .03 and .04 when *d-s* parameters = .0, between .07 and .11 when *d-s* parameters = .15, and between .10 and .15 when *d-s* parameters = .3; but the 95% confidence interval of the average bias for slipping parameters was between .05 and .09 when *c-g* parameters = .0, between .06 and .10 when *c-g* parameters = .15, and between .08 and .12 when *c-g* parameters = .3; see Figure 4a and 4b).

DISCUSSION and CONCLUSION

Multiple-choice test items might be regarded as a popular item type in educational and psychological assessments. However, in a test with multiple-choice test items, some test takers may guess a correct answer (i.e., guessing effect), or miss it because of anxiety or carelessness (i.e., slipping effect). The estimation of item parameters and test-takers' abilities might be biased when the guessing effect and/or the slipping effect are not modeled in data analyses. The DINA model and 4PL IRT model consider the guessing and slipping effects through including a parameter for the guessing effect (i.e., *g* parameter in DINA model and *c* parameter in 4PL IRT model) and a parameter for the slipping effect (i.e., *s* parameter in DINA model and *d* parameter in 4PL IRT model) when analyzing data and estimating model parameters such as item parameters and test-takers' abilities. The current simulation study purported to compare the estimated *c-g* and *d-s* parameters from the 4PL IRT model and DINA model through manipulating the number of attributes, the level of correlation among attributes, test length, the level of *g* parameter, and the level of *s* parameter.

The research findings indicate that the guessing and slipping parameters were estimated correctly across all study conditions when the DINA model was used to analyze the datasets in the study (e.g., the RMSEs of the guessing and slipping parameters were almost zero across all study conditions). The good performance of the DINA model is consistent with the results in the literature (e.g., Chiu, 2008; de la Torre et al., 2010; de la Torre & Lee, 2010). However, an important limitation of the current study is the use of the DINA model for data generation. Fitting the correct model (i.e., DINA model) might be a possible reason for the estimation of slipping and guessing parameters correctly. Thus, it might be helpful to use an empirical dataset for the evaluation of guessing and slipping parameters estimated via 4PL IRT and DINA models in a future study.

A typical test length is 15 or 20 to estimate the model parameters accurately in the CDMs, and the model parameters are estimated more accurately via the DINA model as the sample size becomes larger (de la Torre, 2009; de la Torre et al., 2010). In the current study, the test length was fixed as 20 or 40 items, and the sample size was fixed at 3000 in the data generation. The large sample size and the long test length might be other possible reasons for the estimation of slipping and guessing parameters accurately via the DINA model. Future work may consider investigating the impact of a shorter test length (e.g., < 15 or 20) and a smaller sample size (e.g., < 3000) on the accuracy of guessing and slipping parameters estimated via 4PL IRT and DINA models.

Both guessing and slipping parameters were overestimated when the 4PL IRT model was chosen to estimate these two item parameters in lieu of the DINA model. The number of attributes made a significant difference in the overestimation of both guessing and slipping parameters when the 4PL IRT model was fit to the data. The overestimation of the guessing and slipping parameters from the 4PL IRT model became more severe when the number of attributes was greater in the data generation. While the number of attributes became greater for the conditions with the same test length, there were fewer items per attribute. Parameter estimates tend to be more biased for a shorter test (Hulin, Lissak,

& Drasgow, 1982). This might be a possible reason for the overestimation of the guessing and slipping parameters more severely under the conditions with a greater number of attributes given the same test length.

The value of guessing parameters in the data generation was another significant study condition for the estimation of guessing parameters through the 4PL IRT model. The guessing parameters were overestimated more under the conditions with a larger guessing parameter in the data generation. This was not consistent with the results from DeMars' (2007) study where the overestimation was more severe for the conditions with a lower guessing parameter. DeMars fits a unidimensional 3PL IRT model to the datasets that followed a multidimensional 3PL IRT model whereas we analyzed the datasets with the multidimensional factor structure and the slipping effect through fitting a multidimensional 4PL IRT model to the datasets. In addition, due to the small sample size (i.e., 1000), the estimated guessing parameters were biased towards the mean of prior distribution (i.e., .2) in DeMars' study (i.e., the bias = .05, .02, .01, -.01, and -.03 for c = .10, .15, .20, .25, and .30, respectively). However, a relatively larger sample size (i.e., 3000) was used in the current study. These might be some possible reasons for the difference between the findings. Although the average bias of the guessing parameters became larger for the 4PL IRT model under the conditions with a higher slipping parameter, a weaker correlation among attributes, and a shorter test in the data generation, the bias difference was not significant considering the overlap between the 95% confidence intervals. This is consistent with the findings in the literature considering the impact of test length and correlation among attributes (e.g., Hulin et al., 1982; Svetina, Valdivia, Underhill, Dai, & Wang, 2017).

When the slipping parameters were estimated through the 4PL IRT model, the overestimation of slipping parameters was more severe under the conditions with a greater slipping parameter in the data generation. However, the bias of the slipping parameters from the 4PL IRT model did not differ across the different levels of the guessing parameters, the correlation among attributes, and the test length in the data generation when the 95% confidence interval of the average bias was taken into consideration. The findings related to the estimated slipping parameters may not be generalized to other study conditions, and there is a need for more studies investigating the parameter recovery in the 4PL IRT model under different study conditions. For example, as mentioned before, the sample size was not manipulated in the current study, and the chosen sample size was limited to 3000 for data generation. However, it is common to use a sample size less than 3000 in literature (see Conway & Huffcutt, 2003; Henson & Roberts, 2006; Jackson, Gillaspy, & Purc-Stephenson, 2009). Although it is recommended that the sample size for running a 3PL model or a DINA model should be larger than 1000 to obtain accurate parameter estimates, there is no rule of thumb for the required sample size of the 4PL IRT model (de la Torre et al., 2010; Hulin et al., 1982). Accordingly, the sample size (e.g., < 3000) might be manipulated in future work to investigate the lower limit for the sample size for running a 4PL IRT model. In addition, it might be helpful to study whether the manipulation of sample size will make a difference in the estimation of slipping and guessing parameters by interacting with the other study conditions such as test length and the correlation among attributes.

Although the estimated slipping and guessing parameters were more biased when datasets were analyzed through the 4PL IRT model than the DINA model, the bias of the estimated slipping and guessing parameters from both 4PL IRT and DINA models were reasonably small in general. Overall, the average bias of both guessing and slipping parameters was smaller than .1 across all study conditions, except the conditions with a high guessing/slipping parameter or a great number of attributes in the data generation. Accordingly, both 4PL IRT and DINA models can be preferred for analyzing the datasets contaminated with guessing and slipping effects. However, it is important to consider the aforementioned limitations of the current simulation study before deciding whether the study results can be generalized to other study settings.

Compliance with Ethical Standards

Funding

This research was supported by Pamukkale University Scientific Research Projects Coordination Unit under code ADEP-2018KRM002-063.

REFERENCES

Barton, M. A., & Lord, F. M. (1981). *An upper asymptote for the three-parameter logistic item-response model* (Research Report 18-21). Princeton, NJ: Educational Testing Service. doi: 10.1002/j.2333-8504.1981.tb01255.x

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1-29. doi: 10.18637/jss.v048.i06

Chiu, C. Y. (2008). *Cluster analysis for cognitive diagnosis: Theory and applications* (Doctoral dissertation). Retrieved from https://www.ideals.illinois.edu/handle/2142/80055

Conway, J. M., & Huffcutt, A. I. (2003). A review and evaluation of exploratory factor analysis practices in organizational research. *Organizational Research Methods*, *6*(2), 147-168. doi: 10.1177/1094428103251541

Culpepper, S. A. (2016). Revisiting the 4-parameter item response model: Bayesian estimation and application. *Psychometrika*, *81*(4), 1142-1163. doi: 10.1007/s11336-015-9477-6

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

DeCarlo, L. T. (2011). On the analysis of fraction subtraction data: The DINA model, classification, latent class sizes, and the Q-matrix. *Applied Psychological Measurement*, *35*(1), 8-26. doi: 10.1177/0146621610377081

de la Torre, J. (2008). An empirically based method of Q-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement*, *45*(4), 343-362. doi: 10.1111/j.1745-3984.2008.00069.x

de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational And Behavioral Statistics*, *34*(1), 115-130. doi: 10.3102/1076998607309474

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*(2), 179-199. doi: 10.1007/s11336-011-9207-7

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrica*, *69*(3), 333-353. doi: 10.1007/BF02295640

de la Torre, J., & Douglas, J. A. (2008). Model evaluation and multiple strategies in cognitive diagnosis: An analysis of fraction subtraction data. *Psychometrika*, *73*(4), 595-624. doi: 10.1007/s11336-008-9063-2

de la Torre, J., Hong, Y., & Deng, W. (2010). Factors affecting the item parameter estimation and classification accuracy of the DINA model. *Journal of Educational Measurement*, *47*(2), 227-249. doi: 10.1111/j.1745-3984.2010.00110.x

de la Torre, J., & Lee, Y. S. (2010). A note on the invariance of the dina model parameters. *Journal of Educational Measurement*, *47*(1), 115-127. doi: 10.1111/j.1745-3984.2009.00102.x

de la Torre, J., & Lee, Y. S. (2013). Evaluating the Wald test for item- level comparison of saturated and reduced models in cognitive diagnosis. *Journal of Educational Measurement*, *50*(4), 355-373. doi: 10.1111/jedm.12022

DeMars, C. E. (2007). "Guessing" parameter estimates for multidimensional item response theory models. *Educational and Psychological Measurement, 67*(3), 433-446. doi: 10.1177/0013164406294778

Doornik, J. A. (2018). *An object-oriented matrix programming language Ox (Version 8.0)* [Computer software]. London: Timberlake Consultants Press.

Finch, H. (2010). Item parameter estimation for the MIRT model: Bias and precision of confirmatory factor analysis-based models. *Applied Psychological Measurement*, *34*(1), 10-26. doi: 10.1177/0146621609336112

Finch, H., Habing, B. T., & Huynh, H. (2003, April). *Comparison of NOHARM and conditional covariance methods of dimensionality assessment*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26*(4), 301-321. doi: 10.1111/j.1745-3984.1989.tb00336.x

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer.

Henson, R., & Douglas, J. (2005). Test construction for cognitive diagnosis. *Applied Psychological Measurement*, *29*(4), 262-277. doi: 10.1177/0146621604272623

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*(3), 393-416. doi: 10.1177/0013164405282485

Hoijtink, H., & Molenaar, I. W. (1997). A multidimensional item response model: Constrained latent class analysis using the Gibbs sampler and posterior predictive checks. *Psychometrika*, *62*(2), 171-189. doi: 10.1007/BF02295273

Huebner, A., & Wang, C. (2011). A note on comparing examinee classification methods for cognitive diagnosis models. *Educational and Psychological Measurement*, *71*(2), 407-419. doi: 10.1177/0013164410388832

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*(3), 249-260. doi: 10.1177/014662168200600301

Jackson, D. L., Gillaspy, J. A., & Purc-Stephenson, R. (2009). Reporting practices in confirmatory factor analysis: an overview and some recommendations. *Psychological Methods, 14*(1), 6-23. doi: 10.1037/a0014694

Junker, B. W. (2001). On the interplay between nonparametric and parametric IRT, with some thoughts about the future. In A. Boomsma, M. A. J. Van Duijn, &T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 274-276). New York, NY: Springer-Verlag.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*(3), 258-272. doi: 10.1177/01466210122032064

Liao, W. W., Ho, R. G., Yen, Y. C., & Cheng, H. C. (2012). The four-parameter logistic item response theory model as a robust method of estimating ability despite aberrant responses. *Social Behavior and Personality: An International Journal*, *40*(10), 1679-1694. doi: 10.2224/sbp.2012.40.10.1679

Loken, E., & Rulison, K. L. (2010). Estimation of a four-parameter item response theory model. *British Journal of Mathematical and Statistical Psychology*, *63*(3), 509-525. doi: 10.1348/000711009X474502

Lorenzo-Seva, U., & Ferrando, P. J. (2006). FACTOR: A computer program to fit the exploratory factor analysis model. *Behavior Research Methods, Instruments, & Computers, 38*(1), 88-91. doi: 10.3758/BF03192753

Lord, F. M. (2012). *Applications of item response theory to practical testing problems*. New Jersey, NJ: Lawrence Erlbaum Associates.

Ma, W., & de la Torre, J. (2020). *GDINA: The generalized DINA model framework*: *R package (Version 2.7.9).* Retrieved from https://CRAN.R-project.org/package=GDINA

Magis, D. (2013). A note on the item information function of the four-parameter logistic model. *Applied Psychological Measurement, 37*(4), 304-315. doi: 10.1177/0146621613475471

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika, 64*, 187-212. doi: 10.1007/BF02294535

Meng, X., Xu, G., Zhang, J., & Tao, J. (2019). Marginalized maximum a posteriori estimation for the four-parameter logistic model under a mixture modelling framework. *British Journal of Mathematical and Statistical Psychology,* Advanced online publication. doi: 10.1111/bmsp.12185

Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus user's guide* (8th ed.). Los Angeles, CA: Muthén & Muthén.

R Core Team. (2017). *R: A language and environment for statistical computing* [Computer Software]. Vienna, Austria: R Foundation for Statistical Computing.

Robitzsch, A., Kiefer, T., George, A. C., & Uenlue, A. (2019). *Package 'CDM'.* Retrieved from https://cran.r-project.org/web/packages/CDM/CDM.pdf

Rowley, G. L., & Traub, R. E. (1977). Formula scoring, number-right scoring, and test-taking strategy. *Journal of Educational Measurement*, *14*(1), 15-22. doi: 10.1111/j.1745-3984.1977.tb00024.x

Rulison, K. L., & Loken, E. (2009). I've fallen and i can't get up: Can high-ability students recover from early mistakes in CAT? *Applied Psychological Measurement*, *33*(2), 83-101. doi: 10.1177/0146621608324023

Svetina, D., Valdivia, A., Underhill, S., Dai, S., & Wang, X. (2017). Parameter recovery in multidimensional item response theory models under complexity and nonnormality. *Applied Psychological Measurement, 41*(7), 530-544. doi: 10.1177/0146621617707507

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of educational measurement, 20*(4), 345-354. doi: 10.1111/j.1745-3984.1983.tb00212.x

Vermunt, J. K., & Magidson, J. (2016). *Upgrade manual for latent GOLD 5.1.* Belmont, MA: Statistical Innovations Inc.

Waller, N. G., & Feuerstahler, L. (2017). Bayesian modal estimation of the four-parameter item response model in real, realistic, and idealized data sets. *Multivariate behavioral research*, *52*(3), 350-370. doi: 10.1080/00273171.2017.1292893

Yakar, L. (2017). *Bilişsel tanı ve çok boyutlu madde tepki kuramı modellerinin karşılıklı uyumlarının incelenmesi* (Doctoral thesis). Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi/

Yen, Y. C., Ho, R. G., Laio, W. W., Chen, L. J., & Kuo, C. C. (2012). An empirical evaluation of the slip correction in the four parameter logistic models with computerized adaptive testing. *Applied Psychological Measurement, 36*(2), 75-87. doi: 10.1177/0146621611432862

Tahmin ve Kaydırma Parametrelerinin Kestiriminde 4PL MTK ve DINA Modellerinin Değerlendirilmesi

Giriş

Psikolojik veya eğitimsel testler genellikle adayların bir davranış örneklemini gözlemlemek için kullanılmaktadır. Bu testlerin birçoğu adayların yetenek veya beceri düzeylerinin ölçülmesine odaklanmaktadır. Bu nedenle bir adayın yeteneğinin, bir maddenin doğru cevaplanmasını nasıl belirlediğinin bilinmesi oldukça önemlidir (Lord, 2012). Genellikle bir başarı testinde gerekli bilgiye sahip adayların maddeyi doğru cevaplamaları, sahip olmayanların ise yanlış cevaplamaları beklenir (Rowley & Traub, 1977). Ancak çoktan seçmeli testlerde bu varsayım her zaman geçerli olmayabilir. Bireyin çoktan seçmeli testlerde verdiği cevaplarda; gerçek yeteneğin yansıması görülebilir, doğru cevaba şans başarısı ile ulaşabilir ya da endişe veya dikkatsizlikten kaynaklı yanlış cevaplar görülebilir (Liao, Ho, Yen, & Cheng, 2012; Yen, Ho, Laio, Chen, & Kuo, 2012). Son iki durumda bireylerin yetenek ve madde parametre kestirimleri yanlı olabilir. Bu durum bazı madde tepki kuramı (IRT) ve bilişsel tanı modelleri (CDMs) tarafından ele alınmaktadır. Şans başarısı-tahmin (Pseudo guessing-guess, *c-g*) ve dikkatsizlik-kaydırma (inattention-slip, *d-s*) parametrelerini ele alan 4 parametreli lojistik (4PL) (Barton & Lord, 1981) model ve DINA (Haertel, 1989; Junker & Sijtsma, 2001) model, bu modellere örnek verilebilir. Bu araştırmanın amacı DINA modele uygun olarak farklı koşullarda üretilen veriler üzerinden 4PL Madde Tepki Kuramı (MTK) ve DINA modelleriyle elde edilen *c-g* ve *d-s* parametrelerini karşılaştırmaktır. Böylece her iki model arasındaki farklılıkların ve benzerliklerin ortaya konulması, *c-g* ve *d-s* doğru parametre kestirimini etkileyen faktörlerin belirlenmesi ve bu parametre tasarımlarına sahip araştırmalara katkıda bulunulması amaçlanmıştır.

Yöntem

Verilerin üretimi ve analizi R yazılımı (R Core team, 2017) ile gerçekleştirilmiştir. Veriler DINA modele uygun olarak üretilmiştir. Bu çalışmadaki koşullar belirlenirken literatürde yer alan çalışmalar dikkate alınmıştır (örn., Chiu, 2008; de la Torre, 2008, 2009, 2011; de la Torre & Douglas, 2004, 2008; de la Torre & Lee, 2010, 2013; de la Torre, Hong, & Deng, 2010; DeMars, 2007; Finch, 2010; Finch, Habing, & Huynh, 2003; Henson & Douglas, 2005; Huebner & Wang, 2011; Meng, Xu, Zhang, & Tao, 2019; Waller & Feuerstahler, 2017). Bu doğrultuda veri üretiminde J = 20 ve J = 40 test uzunlukları dikkate alınmıştır. Bunun yanı sıra .0-.15 (düşük), .15-.30 (orta) ve .30-.45 (yüksek) olmak üzere 3 farklı *g* ve *s* parametre düzeyi belirlenmiştir. Özellikler arası korelasyon düzeyleri r = .2 (düşük), r = .5 (orta), ve r = .8 (yüksek) olarak belirlenmiştir. Modellerden elde edilen parametrelerin doğruluğu için örneklem büyüklüğü N = 3000'e sabitlenmiştir. Ayrıca iki farklı özellik sayısı K = 3 ve K = 5 dikkate alınmıştır. Veri üretiminde dört farklı Q-matris kullanılmıştır (2 test uzunluğu x 2 özellik sayısı). Q-matrislerde yer alan her bir madde bir özellik ile ilişkilendirilmiştir. Q-matrislerde yer alan özellikler ile ilişkili madde sayılarının eşit olmasına dikkat edilmiştir. Araştırma kapsamında toplam 108 koşul (3 *g* düzeyi x 3 *s* düzeyi x 3 korelasyon düzeyi x 2 test uzunluğu x 2 özellik sayısı) test edilmiştir. Her bir koşul için 100 veri seti üretilmiştir. Her bir veri seti çok boyutlu 4PL MTK ve DINA modeller ile analiz edilmiştir. Çok boyutlu 4PL MTK'nın veri analizi için uygunluğunu test

etmek için verilerin faktör yapısı Factor 9.2 yazılımı (Lorenzo-Seva & Ferrando, 2006) ile incelenmiştir. Tetrakorik korelasyona dayalı paralel analizler sonucunda, çok boyutluluk varsayımının kullanılan MTK modeline uygun olduğu doğrulanmıştır. Bu çalışma kapsamı dışında olması nedeniyle üretilen verilerin yerel bağımsızlık varsayımını karşıladığı varsayılarak analizler gerçekleştirilmiştir. DINA model analizleri "CDM" (Robitzsch, Kiefer, George, & Uenlue, 2019) paketi ile gerçekleştirilmiştir. 4PL analizleri için "mirt" (Chalmers, 2012) paketi kullanılmıştır. 4PL MTK ve DINA modellerin *c*-*g* ve *d*-*s* parametre kestirimlerinin doğruluğunun değerlendirilmesinde sapma (bias) ve hata kareler ortalaması karekökü (RMSE) değerleri kullanılmıştır. Sapma ve RMSE değerleri hesaplanırken 4PL MTK'nın tahmin ve kaydırma parametrelerinin DINA modeli ile aynı gerçek değere sahip olduğu varsayılmıştır (geniş bilgi için bkz., Culpepper, 2016; Meng ve diğerleri, 2019). Ortalama sapma ve RMSE değerleri %95 güven aralıkları ile rapor edilmiştir.

Sonuç ve Tartışma

Araştırma kapsamında ulaşılan bulgular, DINA modeli kullanıldığında tahmin (şans başarısı) ve kaydırma parametrelerinin ele alınan tüm çalışma koşullarında doğru bir şekilde kestirildiğini ortaya koymuştur. Tüm çalışma koşulları altında DINA modeli kullanıldığında tahmin ve kaydırma parametrelerinin RMSE değerleri sıfıra yakın bulunmuştur. DINA modelin parametre kestiriminde iyi bir performans sergilemesi literatürdeki diğer çalışma sonuçlarıyla uyumludur (örn., Chiu, 2008; de la Torre & Lee, 2010; de la Torre ve diğerleri, 2010). Ancak, veri üretiminde DINA model kullanılması bu çalışmanın önemli bir sınırlılığıdır. Tahmin ve kaydırma parametrelerinin doğru kestirimi, veriler analiz edilirken doğru model olan DINA modelinin kullanılmasından kaynaklanmış olabilir. Bu nedenle 4PL MTK ve DINA modellerinin tahmin ve kaydırma parametrelerinin kestirimi açısından karşılaştırılması için gelecek çalışmalarda gerçek veri setinin kullanılması önerilmektedir.

CDM modellerinde parametrelerin doğru kestirimi için tipik bir test uzunluğunun 15 ila 20 olduğu ve örneklem büyüklüğü arttıkça DINA modeli kullanılarak yapılan parametre kestirimlerinin daha doğru sonuçlar verdiği bilinmektedir (de la Torre, 2009; de la Torre ve diğerleri, 2010). Bu çalışmada veri üretiminde test uzunluğu 20 ve 40 olarak belirlenmiş ve örneklem büyüklüğü 3000'de sabitlenmiştir. Örneklem büyüklüğünün ve test uzunluklarının yeterli olmasının tahmin ve şans parametrelerinin DINA model kestirim doğruluklarında etkili olduğu düşünülmektedir. Bu nedenle sonraki çalışmalarda test uzunluğunun daha kısa tutulmasının ve düşük örneklem büyüklüklerinin söz konusu sonuçlarda ne gibi değişikliklere neden olacağı incelenebilir.

DINA model yerine 4PL MTK modeli kullanıldığında hem tahmin hem de kaydırma parametresinin gerçek değerlerinden daha büyük kestirimlere neden olduğu belirlenmiştir. Bu durumda özellik sayısının önemli olduğu ve özellik sayısı arttıkça tahmin ve kaydırma parametrelerinin 4PL MTK ile kestirilen değerlerinin gerçek değerlerinden daha da uzaklaştığı bulunmuştur. Test uzunluğu sabit tutularak özellik sayısı artırıldığında her bir özellikle ilişkilendirilmiş madde sayısı azalmaktadır. Bu nedenle daha kısa testlerde parametre kestirimi daha yanlı olmaktadır (Hulin, Lissak, & Drasgow, 1982). Bu doğrultuda test sabit tutulurken özellik sayısının artırılmasının tahmin ve kaydırma parametrelerinde daha yanlı kestirimlere neden olduğu düşünülebilir.

Tahmin parametresinin veri üretimindeki değerinin büyük olması 4PL MTK modeliyle kestirilen tahmin parametresinin daha yanlı olmasına neden olmuştur. Benzer şekilde kaydırma parametresinin veri üretimindeki değerini büyütmek, 4PL MTK modeliyle kestirilen kaydırma parametresinin daha yanlı olmasıyla sonuçlanmıştır. Ancak %95 güven aralıkları dikkate alındığında söz konusu parametrelerin özellikler arası korelasyondan ve test uzunluğundan kayda değer bir şekilde etkilenmediği bulunmuştur. Bu sonuç, test uzunluğu ve özellikler arası korelasyon gibi çalışma özellikleri açısından literatürde bulunan sonuçlarla örtüşmektedir (örn., Hulin ve diğerleri, 1982; Svetina, Valdivia, Underhill, Dai, & Wang, 2017).

Her ne kadar 4PL MTK modeliyle elde edilen tahmin ve kaydırma parametreleri DINA modele kıyasla daha yanlı olsa da, bu kestirimlerdeki yanlılığın genel anlamda önemli olmadığı söylenebilir. Örneğin, tüm çalışma koşulları dikkate alındığında tahmin ve kaydırma parametrelerindeki ortalama yanlılığın

genel olarak .1'den küçük olduğu bulunmuştur. Sadece tahmin ve kaydırma parametrelerinin veri üretimindeki değerlerinin yüksek olduğu koşullar ile özellik sayısının büyük olduğu çalışma koşullarında 4PL MTK modeliyle yapılan kestirimlerin yanlılığı .1'den büyük bulunmuştur. Bu sonuçlar dikkate alındığında araştırmacılar tahmin ve kaydırma etkisine sahip verilerin analizlerinde hem DINA modelini hem de 4PL MTK modelini dikkate alabilirler. Ancak bu sonuçları başka çalışma koşullarına genellemeden önce çalışmanın sınırlılıklarının dikkate alınması oldukça önemlidir.

Yukarıda bahsedilen çalışma sınırlıkları dışında bu çalışmada örneklem büyüklüğünün 3000 olarak sabit tutulması başka bir önemli sınırlıktır. Araştırma kapsamında örneklem büyüklüğü belirlenirken, modellerin doğru parametre kestirimleri sağlamasına yetecek bir örneklem büyüklüğü seçimine dikkat edilmiştir. Ancak literatürde 3000'den daha küçük örneklem büyüklüğü sahip çalışmalara rastlamak oldukça mümkündür (örn., Conway & Huffcutt, 2003; Henson & Roberts, 2006; Jackson, Gillaspy, & Purc-Stephenson, 2009). Bunun yanında 3PL MTK modelini veya DINA modelini kullanmak için gerekli minimum örneklem büyüklüğünün 1000 olması tavsiye edilirken 4PL MTK ile madde parametrelerinin doğru kestiriminde gerekli minimum örneklem büyüklüğüne ilişkin çalışmalara ihtiyaç vardır (de la Torre ve diğerleri, 2010; Hulin ve diğerleri, 1982). Bu doğrultuda gelecek çalışmalarda farklı örneklem büyüklüklerini dikkate alarak 4PL MTK modeli için gerekli minimum örneklem büyüklüğü araştırmanın ve örneklem büyüklüğünün diğer çalışma koşullarıyla etkileşimini incelemenin 4PL MTK ile ilgili literatüre önemli katkılar sağlayacağı düşünülmektedir.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2020; 11(2); 147-162

Rating Performance among Raters of Different Experience Through Multi-Facet Rasch Measurement (MFRM) Model *

Muhamad Firdaus Bin MOHD NOH **

Mohd Effendi Ewan Bin MOHD MATORE ***

Abstract
One's experience can greatly contribute to a diversified rating performance in educational scoring. Heterogeneous ratings can negatively affect examinees' results. The aim of the study is to examine raters' rating performance in assessing oral tests among lower secondary school students using Multi-facet Rasch Measurement (MFRM) model indicated by raters' severity. Respondents are thirty English Language teachers clustered into two groups based on their rating experience in high-stakes assessment. The respondents listened to ten examinees' recorded answers of three oral test items and provided their ratings. Instruments include items, examinees' answers, scoring rubric, and scoring sheet used to appraise examinees' competence in three domains which are vocabulary, grammar, and communicative competence. MFRM analysis showed that raters exhibited diversity in their severity level with chi-square χ^2=2.661. Raters' severity measures ranged from 2.13 to -1.45 logits. Independent *t*-test indicated that there was a significant difference in ratings provided by the inexperienced and the experienced raters, *t*-value = -0.96, *df* = 28, p<0.01. The findings of this study suggest that assessment developers must ensure raters are well versed before they can rate examinees in operational settings gained through assessment practices or rater training. Further research is needed to account for the varying effects of rating experience in other assessment contexts and the effects of interaction between facets on estimates of examinees' measures. The present study provides additional evidence with respect to the role of rating experience in inspiring raters to provide accurate ratings.

Keywords: Rating performance, rater-mediated assessment, Multi-faceted Rasch Measurement model, oral test, rating experience.

INTRODUCTION

Rater-mediated assessment is among the types of ubiquitous assessments in the education system around the world. At a global level, rater-mediated assessment is indispensable in high-stakes assessment to appraise examinees' competence in complex traits such as speaking skill, writing skill, and art in order to screen examinees for essential selections such as university enrolment and job interview. However, the use of raters in assessing examinees' competence within the context of high-stakes assessment brings impact on examinees' final marks (Engelhard & Wind, 2018). This impact, known as the rater effect, is systematically attributed to raters' variability and results in variances in observed ratings (Scullen, Mount & Goff, 2000). Negatively, examinees receive marks deviated far from their actual proficiency in the assessed domains (Myford & Wolfe, 2003).

* This research was fully supported by Universiti Kebangsaan Malaysia under the Dana Penyelidikan FPEND (GG-2019-034).
** Graduate Student, University Kebangsaan Malaysia, Faculty of Education, Selangor-Malaysia, muhamad.firdausi@gmail.com, ORCID ID: 0000-0002-5429-6789
*** Senior Lecturer, University Kebangsaan Malaysia, Faculty of Education, Selangor-Malaysia, effendi@ukm.edu.my, ORCID ID: 0000-0002-6369-8501

To cite this article
Mohd Noh, M. F. & Mohd Matore, M. E. E. (2020). Rating performance among raters of different experience through Multi-Facet Rasch Measurement (MFRM) Model. *Journal of Measurement and Evaluation in Education and Psychology, 11*(2), 147-162. doi: 10.21031/epod.662964

Received: 25.12.2019
Accepted: 09.05.2020

Discussion on rating performance among raters is crucial to ensure that examinees are assessed with fairness and reliability. Rating performance can be indicated through raters' severity. Severity is raters' inclination to severely adhere to assessment procedures and consequently may warrant excellent examinees marks lower than their actual ability (Myford & Wolfe, 2003). On the contrary, leniency is raters' tendency to be lenient and generous in awarding marks more than examinees should receive (Wind, 2018). Raters' failure to control their severity and leniency can contribute to variances in awarded scores, thus negatively affect examinees' results.

Differences in rating performance among raters depend on raters' diverse backgrounds, also known as rater's variability. Rating experience is one of the significant rater variability apart from other factors including gender, age, first language, teaching experience, the amount of training they receive, and language proficiency (Eckes, 2015). Previous research on the effect of raters' rating experience on rating performance has shown contradictory findings. Ahmadi Shirazi (2019) and Alp, Epner, and Pajupuu (2018) found out that observed ratings generated by raters with distinct rating experience were not significantly different. However, Attali (2016), Davis (2016), Huang, Kubelec, Keng, and Hsu (2018), Isaacs and Thomson (2013), and Kim (2015) altogether concurred that raters with distinct rating experience showed significantly different performance.

The literature on rating performances is mostly documenting variability that exists among raters, including their rating experience (Eckes, 2015). Appointment of raters with different rating experience is inevitable as there are always novice raters to replace retired raters. Thus any assessment setting would have a combination of novice and experienced raters. Relative to novice raters, experienced raters may be more impacted by their professionalism and expertise as compared to undesired factors such as familiarity and experience. This situation has resulted in the practice of different judgment levels because some raters rate with generosity, and some raters are stringent in awarding marks to examinees due to their different rating experience. Consequently, examinees are judged with varying levels of severity, and it boils down to the extent to which raters can generate scores within the accepted standard. Empirically, conflicting findings emerged from the literature in terms of how raters' experience has impacted rating quality. Raters of different experiences were reported to show distinct rating quality in some studies (Davis 2016; Huang et al. 2018; Kim 2015), but differences were not observed in other studies (Ahmadi Shirazi, 2019; Isaacs & Thomson, 2013; Şahan & Razı, 2020).

Apart from that, the initiative to evaluate raters' rating quality is usually executed through moderation procedure during which another group of raters reviewing examinees' answer scripts after being marked by the first group of raters. The moderation for writing assessment is carried out by reviewing students' answer scripts, but it is not the case with oral tests as it is a hassle to record examinees' answers. Therefore, the moderation process for oral tests is infeasible; thus, no one can monitor if raters do not rate with irrelevant-construct variance. In other words, raters of oral tests are given full trust to execute the scoring procedure, and the validation of scores they award to examinees solely depends on their professionalism and expertise. It renders examinees' future on raters' performance in providing ratings.

Therefore, the current study contributes to the body of knowledge by confirming the extent to which raters' experience can lead to different rating quality among raters within the context of oral test. This study seeks to investigate the rating performance of oral test raters in terms of their severity levels and responds to the question concerning whether raters of different rating experiences produce significantly different ratings. For such purpose, the study is implemented within the context of assessment executed by lower secondary school teachers through replication of Pentaksiran Tingkatan Tiga (PT3) oral test in Malaysia. The specific research objectives guiding the current study are the following:

1. To identify the severity levels practiced by raters in assessing oral test.
2. To identify the difference in rating performance between experienced and inexperienced raters.

Raters' Rating Performance

Raters are individuals appointed by an authoritative body to mark examinees' answers. Raters are required to attend rater training to be adept in items used in the assessment, rubric, rating scales, rating

procedures, and answer keys. This process aims at preparing raters before they execute the rating process in the operational setting. Raters must be well-trained because the rating process highly depends on their professionalism and comprehension, especially for subjectively scored items (Kang, Rubin & Kermad, 2019).

Subjectively scored items require examinees to construct their answers without being given any answer choices (Haladyna & Rodrigues, 2013), such as essay writing and interview. There are also subjective items scored objectively, for instance, short-response items. A significant difference between the two types of items lies in the freedom warranted to raters while scoring (Albano & Rodrigues, 2018). Objectively scored items are marked with rigidity, and answers that are not provided in the answer keys are not acceptable. However, subjectively scored items are more flexible in accepting answers from examinees even though it is not stated in the answer keys, and raters are given the privilege to use their conscience and expertise in judging examinees' answers.

This situation produces construct-irrelevant variance introduced by raters. It may negatively affect the estimates of examinees' competency measure (Bond & Fox, 2015) because it is impossible for all examinees to be rated by one rater in an operational assessment setting (Jones & Wind, 2018). It is also impractical for all appointed raters to rate all examinees due to time constraints, financial and human resources. Hence, raters' rating performance has captured the attention of many previous researchers, primarily in the area of educational assessment, language assessment, and psychology (Engelhard & Wind, 2018). Rating performance is used interchangeably as 'rater effect,' 'rater accuracy,' and 'rater error.' Notably, this concept refers to the variability existed among raters that hinders them from generating a valid and reliable rating score, which may not purely represent examinees' accurate competence level in the assessed domains (Wu & Tan, 2016).

In analyzing the rating performance of raters, many researchers opt for securitizing severity practiced by raters. Severity is one of the indicators used to identify the extent to which raters succeed in producing quality ratings (Eckes, 2015). This indicator is prominent because raters who are too strict or too lenient may precipitate examinees to be judged with injustice (Myford & Wolfe, 2003). For example, highly proficient examinees may be awarded lower marks if they are rated by strict raters. On the contrary, low proficient examinees may receive higher marks if lenient raters score them.

Findings from previous research have depicted that raters' severity level is different based on how they are grouped and assessment context. Attali (2016) contends that raters' severity level is different when they are clustered according to rating experience. Inexperienced raters used varying degrees of severity as compared to experienced raters, especially before any rater training was given. However, both groups of raters were successful in generating homogeneous ratings after training. Huang et al. (2018) found out that raters showed different levels of severity when they are compared according to their first language within the context of language testing. Recently, Ahmadi Shirazi (2019) assigned raters of writing test to rate using two rating methods (holistic and analytical) and concluded that raters of writing test displayed different levels of severity and leniency. Similarly, Kang, Rubin and Kermad (2019) discovered that raters of different first languages applied conflicting patterns of severity. Native speaker raters usually display a high level of severity, while non-native speaker raters rate with lower severity levels.

However, other research studies reach different conclusions, finding the practice of homogenous ratings among raters regardless of how they are grouped. Koizumi, Okabe and Kashimada (2017) argued that the difference in severity levels exhibited by raters of English language oral test was not significant. Similarly, Weilie (2018), who has tasked teachers and non-teachers to mark examinees' answers in oral storytelling test, concluded that both groups of raters manifested indistinguishable patterns of severities.

Rating Experience

Variability among raters influences their rating performance. Variability with significant impact has been found to include rating experience. It has been identified as a major contributing factor for how raters rate examinees' answer scripts. Hence, a growing body of literature has sought to investigate the extent to which raters' rating experience can leave an impact on the way raters score examinees. However, contradicting findings have emerged from the studies.

Raters were reported to manifest different rating quality when compared based on their amount of rating experience. Experienced raters were able to attain higher inter-rater agreement among them in comparison to beginners (Isaacs & Thomson, 2013) and rate with stability and consistent throughout many rating sessions (Kim, 2015). Novice raters, on the other hand, were found to have difficulties in using the rating scales, produced erratic ratings, and did not understand the rating scales accurately. In contrast, raters with little experience manifested problematic rating patterns, tended to modify ratings but improved a lot after several rating sessions (Kim, 2015). It was further corroborated by Attali (2016), who reported that the correlation of marks between trainee raters and experienced raters were considerably different when the marks were compared within the same group. A comparison of marks within trainee raters suggested that the marks are heterogeneous and have more variance as compared to experienced raters. Such observation was a result of their inability to discriminate between good quality answer scripts and lesser ones. Similarly, Davis (2016) observed inconsistent ratings between experienced and new raters, especially in terms of their severity, reliability, and inter-rater agreement.

On the contrary, other studies have discovered contrasting findings. Alp, Epner, and Pajupuu (2018) concluded that raters with different rating experience managed to achieve acceptable standards of ratings under a condition in which raters were aware of rating procedures. Ahmadi Shirazi (2019), who employed raters with diversified rating experience to mark 20 examinees' answer scripts, reported that raters could rate within an acceptable range of severity level consistently. Raters were also observed to use similar strategies and focused on the same criteria while scoring regardless of their rating experience (Şahan & Razı, 2020).

The contradicting findings that emerged from the literature may be due to the different contexts used in the studies and the research designs employed. Hence, it is indecisive to claim that rating experience is a potent determinant in raters' rating quality. The findings from existing studies also fail to generalize the impact of raters' rating experience. This indicates a need for more research conducted to investigate how their experience can differentiate raters rating quality.

Multi-Faceted Rasch Measurement (MFRM) Model

The multi-faceted Rasch Measurement (MFRM) model is an extension of the Rasch measurement model. The basic of Rasch model allows the calibration of only two estimates, item difficulty and person ability involved in analyzing dichotomous items. MFRM extends the basic logistic dichotomous Rasch model by allowing analysis to include more than two facets of the assessment settings, and the data aimed to be analysed is not necessarily dichotomous (Eckes, 2019). It is therefore probable that additional facets are to be incorporated into the analysis depending on the interest and condition of the assessment. Eckes (2019) elaborated that other facets may include criteria, raters, interlocutors, tasks, and assessment occasions. In order for any study to use MFRM as its primary statistical analysis, the involved facets need to be identified first (Wesolowski & Wind, 2019). After the relevant facets have been presupposed, a suitable MFRM model can formally be expressed to measure the estimation of each facet. MFRM model to calibrate facets in oral tests can be translated into expression as follows:

$$\left(\frac{\mathrm{P}^{n\mathrm{ljmk}}}{\mathrm{P}^{n\mathrm{ljmk}}-1}\right) = \theta^{n} - \delta^{\mathrm{l}} - \alpha^{\mathrm{j}} - \upsilon^{\mathrm{m}} - \tau^{\mathrm{k}} \quad (1)$$

where

P^{nljmk} = probability of examinee *n* receiving a rating of *k* from rater *j* on domain *m* for item *l*

$P^{nljmk-1}$ = probability of examinee *n* receiving a rating of *k-1* from rater *j* on domain *m* for item *l*

θ_n = ability of examinee *n*,

δ_l = difficulty of item *l*,

α_j = severity of rater *j*,

υ_m = difficulty of domain *m*,

τ_k = difficulty of receiving a rating of *k* relative to *k-1*

Based on the four-facet MFRM model shown in Equation 1, MFRM is an additive-linear model that enables observed ratings to be transformed into a logit scale (Myers, Well & Lorch, 2010). The estimation of each facet will be calibrated using the logit scale. MFRM yields analysis of raters with several statistics, including estimation of measures for each measure presented in a graphical Wright map, separation statistics, fit statistics, and also inter-rater agreement (Eckes, 2015).

METHOD

Research Design

This quantitative study through survey design was executed by simulating English Language oral test for lower secondary school students. The survey enables the study to be implemented using a small number of respondents, and data can be collected with minimal financial support and within a short period of time (Creswell & Creswell, 2018).

Respondents

A total of 30 lower secondary school English teachers in the state of Selangor were involved as respondents in this study. Selangor was chosen because it has the highest number of teachers (Kementerian Pendidikan Malaysia, 2019a), resulting in a heterogeneous background among teachers as compared to other states. Meanwhile, English was selected because it is a tough subject for Malaysian students sitting for public examinations compared to other subjects (Kementerian Pendidikan Malaysia, 2019b). Thus, teachers' competence to appraise students' proficiency in English needs absolute attention. The respondents were divided into two groups based on their experience in rating high-stakes assessment, especially PT3. The first group (Rater 1 to Rater 15) consists of teachers who do not have any experience in rating high-stakes assessments other than carrying out assessment only in the classroom. The second group (Rater 16 to Rater 30) are experienced teachers with a minimum of two years of experience in rating high-stakes assessment.

Instrumentation

Instruments used in the study were items for oral test, examinees' recorded answers, scoring rubric, and scoring form. Questions were adapted from an oral test exercise book (Anthony & Miriam, 2019). Three oral test items were used, which include background interview, storytelling based on pictures, and a discussion based on a mind map. Ten lower secondary school students of mixed proficiency levels were chosen to answer the questions by simulating the actual assessment scenes like in PT3. An English teacher who is experienced in conducting the PT3 oral test was appointed as an interlocutor to carry out the test. The students' answers were recorded using a recorder.

The scoring rubric was adapted from lower secondary school (form one, two, and three) oral tests rubric in the *Common European Framework Reference for Language* (CEFR) 2019 established by the ministry (Lembaga Peperiksaan, 2019). Three domains were assessed, vocabulary (Domain 1), grammar (Domain 2), and communicative competence (Domain 3). Each domain is divided into five different mastery levels, which are level 1 (the lowest), level 2 (low), level 3 (average), level 4 (high), and level 5 (the highest). The scoring sheet is used by raters to record each examinee's mark. All the instruments have undergone face and content validity procedures involving nine-panel of experts. These panels are university lecturers who are experts in language testing and educational measurement. Inter-rater agreement was fully achieved, and their qualitative comments were considered before the instruments were used in collecting data.

Administration

The rating process was implemented by all raters who were assigned to rate all examinees' answers. It was done using a fully-crossed rating design to ensure connectedness among presupposed facets (Engelhard & Wind, 2018), as shown in Table 1. This design was used by previous research to create sufficient linkage and enable rating performance analysis (Wind & Sebok-Syer, 2019). Each rater was required to listen to the recordings and give ratings for item one involving domain one and two, item two involving domain one and two and also item three involving domain one, two and three as summarised in Table 1. Altogether, each rater has generated 70 scores (domain 1,2,1,2,1,2,3 x ten examinees).

Table 1. Assessment Mapping Implemented by Raters

Raters	Items	Domains	Examinees' answer recordings									
			1	2	3	4	5	6	7	8	9	10
Rater 1	1	1,2	√	√	√	√	√	√	√	√	√	√
↓	2	1,2	√	√	√	√	√	√	√	√	√	√
	3	1,2,3	√	√	√	√	√	√	√	√	√	√
	1	1,2	√	√	√	√	√	√	√	√	√	√
	2	1,2	√	√	√	√	√	√	√	√	√	√
	3	1,2,3	√	√	√	√	√	√	√	√	√	√
	1	1,2	√	√	√	√	√	√	√	√	√	√
	2	1,2	√	√	√	√	√	√	√	√	√	√
	3	1,2,3	√	√	√	√	√	√	√	√	√	√
	1	1,2	√	√	√	√	√	√	√	√	√	√
Rater 30	2	1,2	√	√	√	√	√	√	√	√	√	√
	3	1,2,3	√	√	√	√	√	√	√	√	√	√

Statistical Analysis

In total, the number of ratings generated by all the raters was 2,100. The data was then analyzed using MFRM model through FACETS software version 46.7.1 (Linacre, 2014a). This software can calibrate more than two facets on the interval logit scale. The software is not only able to identify the interaction between item difficulty and examinees' ability but also raters' severity by producing Wright map, separation statistics, and fit statistics (Linacre, 2014b). MFRM is used because of its suitability, and researchers of rating performance have frequently employed this approach to investigate rater effects either in simulation or real-data studies (Wind & Guo, 2019).

The assumption of the Rasch model was met in terms of item fit and is depicted in Table 2. The findings have revealed that the *infit* MNSQ of all the three items used was ranged between 0.91 to 1.05, and the range for the *outfit* was between 0.87 to 1.07. Meanwhile, the Zstd values were reported to be within ±2.0 range as recommended by Bond and Fox (2015) except for one item, Storytelling (2.1). The standard error which indicates the precision of measurement (Linacre, 2005) for all the items was ranged between 0.7 to 0.9. The range of standard error is classified as excellent since they are under 0.25 (Fisher,

2007). As for the PTMEA, positive values of more than 0.30 are desirable (Wu & Adam, 2007). All three items managed to achieve the desired value ranged from 0.77 to 0.85. The PTMEA values indicate that the items were able to discriminate the abilities of the candidates in assessing their speaking skills. Overall, all three items were fit and suitable to be used in the study.

Table 2. Item Fit Report

Items	Measure	Model S.E.	Infit		Outfit		Estim. Discrm	Correlation	
			MnSq	ZStd	MnSq	ZStd		PtMea	PtExp
Interview	-1.84	0.09	1.05	0.8	1.07	1	0.95	0.77	0.82
Storytelling	1.1	0.09	0.91	-1.6	0.87	-2.1	1.09	0.81	0.81
Discussion	0.74	0.07	1.01	0.1	1.02	0.4	0.98	0.85	0.82
Mean	0	0.08	0.99	-0.2	0.99	-0.2	-	0.81	-
SD Population	1.31	0.01	0.06	1	0.09	1.4	-	0.04	-
SD Mean	1.61	0.01	0.07	1.3	0.11	1.7	-	0.04	-

To determine the functioning of each response category, Linacre's (2002) guidelines for evaluating rating scale category effectiveness were applied to the data. Table 3 shows the statistical report of the scales used in the study.

Table 3. Scale Report

Data			Quality Control		Outfit MnSq	Rasch-Andrich Threshold		Exp. Meas. at		Most Probable for	Rasch-Thurstone Threshold	Cat Peak Prob
Category Score	Used	%	Avrge. Meas.	Exp. Meas.		Meas.	S.E.	Category -0.5				
1	145	7	-6.71	-6.95	1.2			-7.95		low	low	100%
2	710	34	-3.65	-3.56	1	-6.88	0.12	-4.43	-6.88	-6.88	-6.89	85%
3	865	41	-0.2	-0.23	0.9	-2.07	0.07	0.01	-2.07	-2.07	-2.08	80%
4	363	17	2.6	2.58	1	2.11	0.08	4.44	2.11	2.11	2.1	84%
5	16	1	4.51	4.73	1	6.85	0.27	-7.93	6.86	6.85	6.84	100%

For any rating scale to be considered of high quality, Linacre advocated six basic conditions to be met. Firstly, a minimum of ten observations for each category was evident as the use of each category score was ranged between 16 to 865. Secondly, average category measures that increase monotonically with categories were observed as the average measures have increased in an orderly manner from -6.71 to -3.65 to 0.2 to 2.6 to 4.51. Thirdly, *outfit* mean square statistics less than 2.0 was attained as the values of all the category scores were ranged between 0.9 to 1.2. Fourthly, Rasch-Andrich category thresholds that increase monotonically was fulfilled as the values have increased from -6.88 to -2.07 to 2.11 to 6.85. Fifthly, Rasch-Andrich category thresholds should be 1.0 to 5.0 logits apart. As shown on Table 4, the threshold between the scale categories in this study ranged between 1.0 to 5.0 except for Scale 1 and Scale 2 with difference value, 6.88. Finally, it was also observed that the shape of the probability curves peaked for each category as presented in Figure 1. The peaks of all the category scores can be clearly seen. Therefore, all five scales were appropriate to be used in the study.

Table 4. Threshold Change (gaps)

Pair of scale	Gaps	Threshold results
S_{1-2}	0.00 – (- 6.88)	6.88 (> 1.0)
S_{2-3}	- 6.88 – (- 2.07)	4.81(> 1.0)
S_{3-4}	- 2.07 – (2.11)	4.18 (> 1.0)
S_{4-5}	2.11 – (6.85)	4.74 (> 1.0)

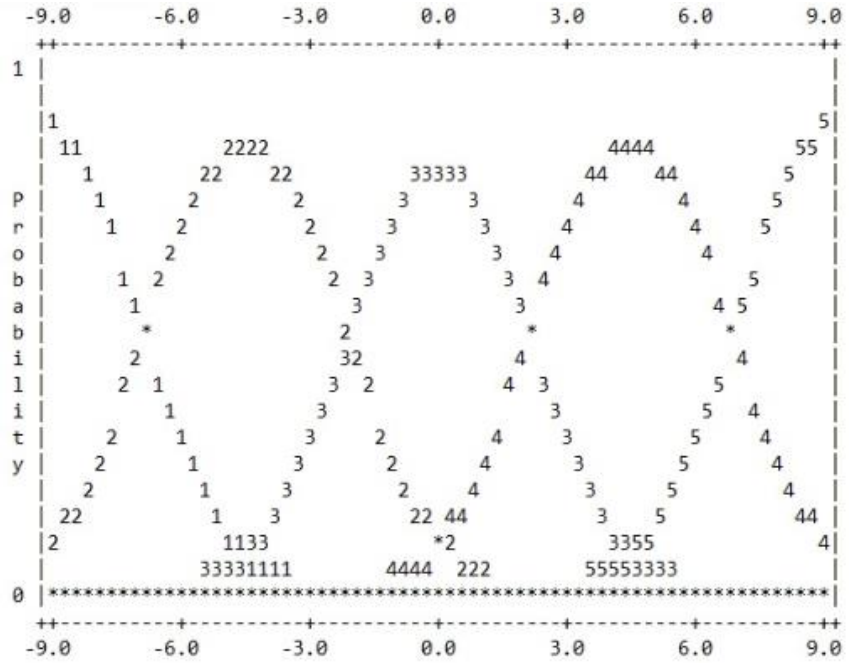


Figure 1. Threshold for scale review

RESULTS

Four facets examined in this study were examinees, items, raters and domains. In addition, raters' rating experience was included as a dummy facet only and not to recognize its effect on estimation of other facets but merely to see the difference of ratings generated by raters of different experience. Figure 2 presents Wright map, a graphical summary of the estimates of all facets. The first column is interval-logit scale used to calibrate all the other facets. The second column compares the ten examinees in terms of their ability in the oral test starting from the most able examinee at the top to the least able examinee at the bottom of the column. Next, the third column compares all the raters based on their severity level. The most severe rater is located at the top and the most lenient rater is positioned at the bottom. The fourth column shows the three items used in the oral test based on difficulty level. The fifth column displays domains assessed in the test arranged based on their difficulty levels starting from the most difficult at the top and the least difficult at the bottom.

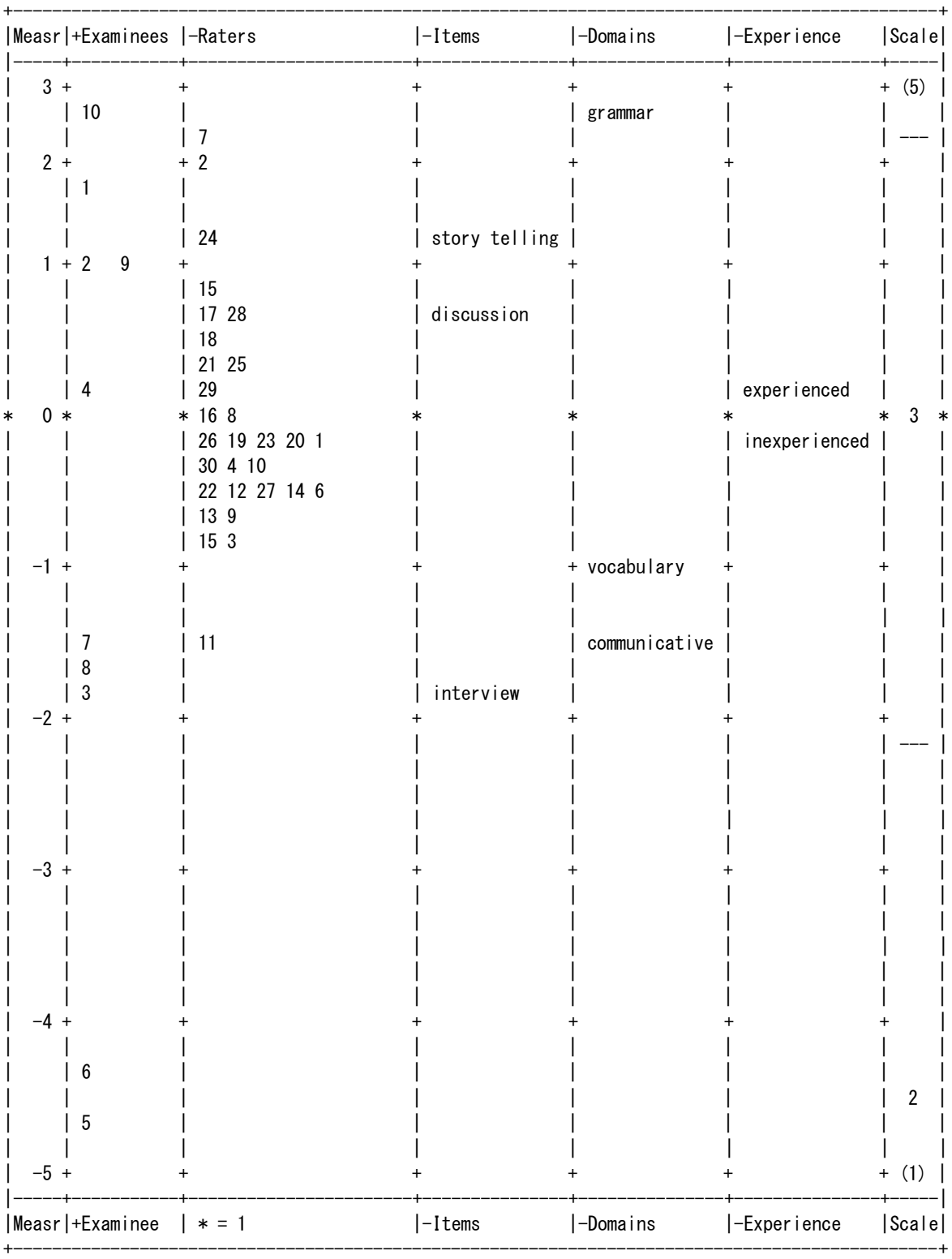


Figure 2. Wright Map of Examinees, Raters, Items, and Domains

Table 5 summarizes MFRM statistics for examinees, raters, items, and domains in terms of their mean, standard error, infit, outfit, chi-square value, and separation statistics. The separation statistics provide

separation ratio, separation index, and separation reliability. However, only rater facet is further analyzed as this study only aims at scrutinizing rating performance among raters.

Table 5. Summary of MFRM Statistics

Statistics	Examinees	Raters	Items	Domains
M Measure	-0.78	0.00	0.00	0.00
M SE	0.15	0.25	0.08	0.09
Infit	1.00	0.99	0.99	1.01
Outfit	0.99	0.99	0.99	1.03
χ^2	2348.5	266.2	718.6	1503.4
df	9	29	2	2
Separation Ratio	15.67	2.85	15.98	19.45
Separation Index	21.22	4.13	21.64	26.27
Separation Reliability	1.00	0.89	1.00	1.00

Objective 1: To Identify the Severity Level Practiced by Raters in Assessing Speaking Test

Based on Table 5, the analysis of chi-square for the homogeneity test indicates that the severity of at least two raters was heterogeneous, with chi-square value $\chi^2 = 266.2$, $df = 29$, $p<.01$. Therefore, the null hypothesis saying that there was no difference in severity practiced by all the raters was rejected. The rater separation ratio intends to inform the spread of the facet measures relative to the precision of those measures (Govindasamy, Salazar, Lerner & Green,2019). The rater separation ratio is 2.85, suggesting that the difference of severity among raters was almost three times than measurement error. The separation strata index is meant to statistically quantify how many different classes of rater, which ideally should be close to 1 if the raters are required to exhibit identical severity patterns (Eckes, 2019). The separation index for the current study is 4.13 indicating there were more than four statistically different strata of rater severity that emerged from the 30 raters. Briefly, the raters did not make a homogenous group, and even the mean standard error was also small, only at 0.25. The next separation statistics is separation reliability, which indicates the overall precision of rater severity estimates and the extent to which differences among raters are measured according to the correct measurement procedures (Wesolowski & Wind, 2019). The reliability of separation statistic in this study is high, 0.89 suggesting that the rater severity variance appeared from the analysis was precise and not affected by measurement errors.

The Wright map shown in Figure 2 presents logits value for rater measure ranged between 2.13 (Rater 7) to -1.45 (Rater 11). Even though there was severity difference observed among the raters, the differences were not that distant because 26 raters were located within 1.0 to -1.0 logit. Eckes (2019) proposes that raters with severity estimates ≥ 1.0 logits are classified as "severe raters" and raters with severity estimates ≤-1.0 logits are "lenient raters." In this study, there were only three severe raters, Rater 11 (2.13 logits), Rater 2 (1.92 logits), and Rater 24 (1.12 logits) and only one lenient rater, that is Rater 11 (-1.45 logits). Such observation was a result of raters' varying abilities in understanding the scoring rubric well enough and their familiarity in assessing speaking skills that was gained through assessment routines carried out in classroom-based context or high-stakes assessments (Kang, Rubin & Kermad, 2019).

Next, further analysis is needed through fit statistics of raters specifically because the measures of raters were proven heterogeneous. Fit statistics in MFRM are used to indicate how raters are consistent in using the rating scales across examinees, items, and domains (Eckes, 2019). Additionally, the statistics also inform the degree to which raters are consistent in arranging examinees according to their ability (Engelhard & Wind, 2018). It also functions to determine the extent to which the ratings generated by raters match what is expected by the measurement model (Wesolowski & Wind, 2019) by analyzing any gap between the observed scores and the expected scores (Wu, 2017). Mean square (MNSQ) of infit and outfit statistics are commonly used to determine the location of raters and other facets (Eckes, 2019). Infit MNSQ indices are functional in identifying inliers' fit (Wu & Tan, 2016). The acceptable range for fit statistics is within 0.50 to 1.50 (Linacre, 2002). There are two indices in fit statistics, misfit and

overfit. Fit statistic less than 0.5 is considered overfit, or raters do not exhibit enough variations in their ratings, while fit statistics greater than 1.5 indicates misfit or too much unpredictability (Wu & Tan, 2016). Eckes (2015) warned that misfit raters are more problematic than overfit raters.

Based on the infit statistics displayed in Table 6, there was only one misfit rater, Rater 13 (with infit MNSQ value 1.55). It implies that Rater 13 exhibited inconsistent rating patterns throughout the rating session. It is interesting to note that this rater was from the inexperienced rater group. This finding conforms to Weilie (2018), who spotted one misfit rater among non-teacher raters that did not have any experience related to rating work. However, surprisingly, Ahmadi Shirazi (2019), who assigned raters to rate using holistic scoring, found that two misfit raters were those with more than five years of rating experience. On top of that, Isaacs and Thomson (2013) figured that there was no clear pattern for misfit raters based on their rating experience because the findings revealed that from eleven misfit raters, five were experienced raters while six were novice raters. Briefly, these results suggest that misfit occurrence was not necessarily due to raters' rating experience. In fact, the other 14 inexperienced raters in this study were located within the acceptable range of infit statistics. In addition, there was no case of overfitting raters as none of the raters were indicated with logits measure less than 0.50. The absence of overfitting occurrence means that no raters produced ratings that were too consistent or easily could be predicted (Jeong, 2017).

Table 6. MFRM Summary of Rater Facet

Raters	Severity logits	Infit MNSQ	Raters	Severity logits	Infit MNSQ
7	2.13	1.15	20	-0.19	0.57
2	1.92	1.16	1	-0.24	1.02
24	1.12	1.50	30	-0.25	1.44
15	0.79	0.99	4	-0.30	1.28
17	0.72	0.90	10	-0.36	0.64
28	0.65	1.23	22	-0.45	0.53
18	0.52	0.59	12	-0.49	1.40
21	0.33	1.18	27	-0.50	1.10
25	0.26	0.68	14	-0.55	0.85
29	0.20	0.63	6	-0.55	1.36
16	0.07	1.47	13	-0.68	1.55
8	0.02	1.00	9	-0.68	0.72
26	-0.12	0.60	5	-0.81	0.69
19	-0.12	0.68	3	-0.81	0.93
23	-0.19	0.75	11	-1.45	1.15

Inter-rater agreement opportunities: 14687; Exact agreements: 9468 = 64.5%; Expected: 8235.7 = 56.1%

Next, MFRM also highlighted inter-rater agreement among raters by comparing it to what the measurement model has suggested. Inter-rater agreement advocates the correlation of marks assigned by all raters (Wu & Tan, 2016). The raters in this study managed to attain 64.5% of inter-rater agreement, higher than what the model has expected, which was 56.1%. It infers that all the raters were able to provide ratings that were beyond the acceptable threshold of inter-rater agreement expected by the model. This convergence may indicate that most raters were able to interpret the scoring rubrics in a similar way (Wu & Tan, 2016).

Objective 2: To Identify the Difference in Rating Performance Between Experienced and Inexperienced Raters

Raters were divided into two groups based on their rating experience. Severity indicator is then compared to examine the difference in severity for both groups exhibited through independent sample *t*-test. Table 7 presents the mean logits and standard deviation for both the inexperienced rater group (M

= -0.14, SD = 1.00) and the experienced rater group (M = 0.14, SD = 0.47). The mean logits show that the severity level of both groups did not deviate far from the total mean logits positioned at 0 logits.

Table 7. Differences of Rater Severity Based on Experience

Groups	N	Mean logits	Standard Deviation	Standard Error	Inter-rater agreement
Inexperienced raters	15	-0.14	1.00	0.26	58.9%
Experienced raters	15	0.14	0.47	0.12	70.0%

t value = -0.96; *df* = 28; *p*<0.01

The analysis of the independent sample *t*-test indicates that there was a statistically significant difference between the two groups of raters with *t*-value = -0.96, *df* = 28, *p*<0.01. It means that the null hypothesis that there was no difference between ratings provided by the inexperienced and the experienced raters was rejected. It signifies that the severity practiced by the two groups was not identical. This finding is consistent with those of Attali (2016); Davis (2016) and Huang et al. (2018), who reported that raters with varying rating experience provided heterogeneous ratings, even though the studies were implemented in different contexts. This consistency may be due to how rating experience among raters was operationally defined. Raters in the aforementioned studies, including the current study, were categorized based on whether they have rating experience in high-stakes assessment or not.

Furthermore, the two groups of raters differed in terms of inter-rater agreement. The experienced raters were able to attain 70.0% inter-rater agreement, while the inexperienced raters only managed to achieve 58.9% inter-rater agreement. This finding is in agreement with Isaacs and Thomson's (2013) findings, which showed that inter-rater agreement among experienced raters was higher than among inexperienced raters. It may be the case, therefore, that experienced raters managed to rate with a mutual understanding of rubric and procedures. Indeed, it is desirable that raters manage to yield quality ratings, especially in terms of inter-rater reliability, despite their variability.

CONCLUSION and DISCUSSION

The present study was designed to determine rating performance between inexperienced and experienced raters within the context of oral tests in addition to confirming findings observed from previous studies despite being conducted in different contexts. Through the analysis of MFRM, one of the significant findings emerged from this study was that raters with different experiences showed non-uniform severity level whereas, the experienced raters displayed more consistency than the inexperienced raters. In general, therefore, the findings indicate that rating experience plays an important role in determining the quality of ratings provided by raters. It is important to note especially by assessment developers that raters with different rating experiences may produce distinct rating quality. Since it is inevitable to avoid the appointment of new raters to replace retired raters, it is noteworthy to ensure that raters undergo sufficient training sessions before engaging in operational assessment routines. Additionally, training for raters must incorporate enough practical scoring opportunities by simulating real situations of assessment conditions so that they can increase their ability to rate examinees. A number of caveats need to be noted regarding the present study. While the study was based on small sample size, the study was also carried out only within lower secondary school oral test practicea. Research is also needed to determine how findings will be different if tested on broader samples and contextualized in other assessment settings. Apart from that, this study has only discussed the rater facet even though analysis of other facets (examinees, items and domains) were also generated by MFRM. In fact, the rater facet was only analysed using the severity indicator. It would be interesting to compare raters' rating performance using other indicators such as halo effect and central tendency. Future studies can also examine the effects of interaction between facets on the estimates of examinees' measures. Additionally, it is unfortunate that the study did not include any rater training prior to scoring sessions. Therefore, it is recommended that further research to include rater training before raters are

engaged in scoring procedures so that the effects of training can be clearly identified between raters with distinct rating experience.

ACKNOWLEDGEMENT

This research was fully supported by Universiti Kebangsaan Malaysia under the Dana Penyelidikan FPEND (GG-2019-034). Appreciation is also given to all respondents involved in answering the survey. Thanks to all the expert panels for assistance with instrument validation, data analysis and initial comments on the manuscript.

REFERENCES

Ahmadi Shirazi, M. (2019). For a greater good: Bias analysis in writing assessment. *SAGE Open*, *9*(1), 1–14. https://doi.org/10.1177/2158244018822377

Albano, A. D., & Rodrigues, M. (2018). Item development research and practice. *Handbook of Accessible Instruction and Testing Practices: Issues, Innovations, and Applications*, 181–198. https://doi.org/10.1007/978-3-319-71126-3_12

Alp, P., Epner, A., & Pajupuu, H. (2018). The influence of rater empathy, age and experience on writing performance assessment. *Linguistics Beyond And Within*, *3*(2017), 7–19. Retrieved from https://www.ceeol.com/search/article-detail?id=716601

Anthony, L., & Miriam, S. (2019). *Drill in English Skills Practice: CEFR-Alligned Curriculum*. Selangor: Oxford Fajar

Attali, Y. (2016). A comparison of newly-trained and experienced raters on a standardized writing assessment. *Language Testing*, *33*(1), 99–115. https://doi.org/10.1177/0265532215582283

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch Model Fundamental Measurement in the Human Sciences*. New Jersey: Lawrence Erlbaum Associates.

Creswell, J. W., & Creswell, J. D. (2018). *Research Design: Qualitative, Quantitative, And Mixed Methods Approaches* (5th ed.). California: Sage Publications.

Davis, L. (2016). The influence of training and experience on rater performance in scoring spoken language. *Language Testing*, *33*(1), 117–135. https://doi.org/10.1177/0265532215582282

Eckes, T. (2015). *Introduction to Many-Facet Rasch Measurement: Analyzing and Evaluating Rater-Mediated Assessments* (2nd ed.). Peter Lang.

Eckes, T. (2019). Implications for rater-mediated language assessment. In Aryadoust, V., & Raquel, M. (Eds.), *Quantitative Data Analysis for Language Assessment Volume I: Fundamental Techniques* (pp. 153-175). London & New York: Routledge.

Engelhard, G., & Wind, S. A. (2018). *Invariant Measurement with Raters and Rating Scales: Rasch Models for Rater-Mediated Assessments*. *Routledge*. New York & London: Routledge. https://doi.org/10.1017/CBO9781107415324.004

Fisher, J. W. P. 2007. Rating scale instrument quality criteria. *Rasch Measurement Transactions* 21(1): 1095.

Govindasamy, P., Salazar, M. D. C., Lerner, J., & Green, K. E. (2019). Assessing the reliability of the framework for equitable and effective teaching with the many-facet rasch model. *Frontiers in Psychology*, *10*(June), 1–10. https://doi.org/10.3389/fpsyg.2019.01363

Haladyna, T. M., & Rodrigues, M. C. (2013). *Developing and Validating Test*. New York: Routledge.

Huang, L., Kubelec, S., Keng, N., & Hsu, L. (2018). Evaluating CEFR rater performance through the analysis of spoken learner corpora. *Language Testing in Asia*, *8*(1), 1–17. https://doi.org/http://dx.doi.org/10.1186/s40468-018-0069-0

Isaacs, T., & Thomson, R. I. (2013). Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions. *Language Assessment Quarterly*, *10*(2), 135–159. https://doi.org/10.1080/15434303.2013.769545

Jeong, H. (2017). Narrative and expository genre effects on students, raters, and performance criteria. *Assessing Writing*, *31*, 113–125. https://doi.org/10.1016/j.asw.2016.08.006

Jones, E., & Wind, S. A. (2018). Using Repeated Ratings to Improve Measurement Precision in Incomplete Rating Designs. *Journal of Applied Measurement*, *19*(2), 148–161. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/29894984

Kang, O., Rubin, D., & Kermad, A. (2019). The effect of training and rater differences on oral proficiency assessment. *Language Testing*, *36*(4), 481–504. https://doi.org/10.1177/0265532219849522

Kementerian Pendidikan Malaysia. (2019a). *Quick Facts 2018: Malaysia Education Statistics.* Retrieved from https://www.moe.gov.my/en/muat-turun/laporan-dan-statistik/quick-facts-malaysia-education-statistics/563-quick-facts-2018-malaysia-educational-statistics/file

Kementerian Pendidikan Malaysia. (2019b). *Pengumuman Analisis Keputusan Sijil Pelajaran Malaysia (SPM) 2018.* Retrieved from http://lp.moe.gov.my/images/bahan/spm/2019/14032019/Laporan%20Analisis%20Keputusan%20SPM%202018%20-%20Upload.pdf

Kim, H. J. (2015). A qualitative analysis of rater behavior on an L2 speaking assessment. *Language Assessment Quarterly*, *12*(3), 239–261. https://doi.org/10.1080/15434303.2015.1049353

Koizumi, R., Okabe, Y., & Kashimada, Y. (2017). A Multi-faceted Rasch analysis of rater reliability of the Speaking section of the GTEC CBT. *ARELE: Annual Review of English Language Education in Japan*, 241–256. Retrieved from https://www.jstage.jst.go.jp/article/arele/28/0/28_241/_article/-char/ja/

Lembaga Peperiksaan. (2019). Instruction to Speaking Examiners (Pentaksiran Tingkatan 3). Retrived from http://lp.moe.gov.my/images/bahan/pt3/2019/21082019/S1%20MES%20PT3%20Instructions%20to%20Speaking%20%20Examiners_Revised%20version.pdf

Linacre, J. M. (2005). Standard errors: means, measures, origins and anchor values. *Rasch Measurement Transactions*, *19*(3), 1030.

Linacre J. M. (2002). Understanding Rasch measurement: Optimizing rating scale category effectiveness. *Journal of Applied Measurement*, *3*, 85–106. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.424.2811&rep=rep1&type=pdf

Linacre, J. M. (2014a). Facets Rasch measurement computer program (Version 3.71.4) [Computer software]. Chicago: Winsteps.com.

Linacre, J. M. (2014b). A user's guide to FACETS: Rasch-model computer programs. Chicago: Winsteps.com. Retrieved from http://www.winsteps.com/facets.htm

Myers, J. L., Well, A. D., & Lorch, R. F. (2010). *Research design and statistical analysis* (3rd ed.). New York, NY: Routledge

Myford, C., & Wolfe, E. W. (2003). Detecting and measuring rater effects using Many-Facet Rasch Measurement: Part I. *Journal Od Applied Measurement*, *4*(October 2015), 386–422. Retrieved from https://www.researchgate.net/profile/Carol_Myford/publication/9069043_Detecting_and_Measuring_Rater_Effects_Using_Many-Facet_Rasch_Measurement_Part_I/links/54cba70e0cf298d6565848ee.pdf

Şahan, Ö., & Razı, S. (2020). Do experience and text quality matter for raters' decision-making behaviors? *Language Testing*, 1-22. https://doi.org/10.1177/0265532219900228

Scullen, S. E., Mount, M. K., & Goff, M. (2000). Understanding the latent structure of job performance ratings. *Journal of Applied Psychology*, *85*(6), 956–970. https://doi.org/10.1037/0021-9010.85.6.956

Weilie, L. (2018). To what extent do non-teacher raters differ from teacher raters on assessing story-retelling. *Journal of Language Testing & Assessment*, *1*, 1–13. Retrieved from http://clausiuspress.com/assets/default/article/2018/08/29/article_1535590233.pdf

Wesolowski, B. C., & Wind, S. A. (2019). Pedagogical considerations for examining rater variability in rater-mediated assessments: A three- model framework. *Journal of Educational Measurement*, *56*(3), 521–546. https://doi.org/10.1111/jedm.12224

Wind, S. A., & Sebok-Syer, S. S. (2019). Examining differential rater functioning using a between-subgroup outfit approach. *Journal of Educational Measurement*, *56*(2), 217–250. https://doi.org/10.1111/jedm.12198

Wind, S. A. (2018). Examining the impacts of rater effects in performance assessments. *Applied Psychological Measurement*, *43*(2), 159–171. https://doi.org/10.1177/0146621618789391

Wind, S. A., & Guo, W. (2019). Exploring the combined effects of rater misfit and differential rater functioning in performance assessments. *Educational and Psychological Measurement*, 1–26. https://doi.org/10.1177/0013164419834613

Wu, M. (2017). Some IRT-based analyses for interpreting rater effects. *Psychological Test and Assessment Modeling*, *59*(4), 453–470. Retrieved from https://www.psychologie-aktuell.com/fileadmin/download/ptam/4-2017_20171218/04_Wu.pdf

Wu, M. & Adams, R. (2007). A*pplying the Rasch model to psycho-social measurement: A practical approach.* Melbourne: Educational Measurement Solutions

Wu, S. M., & Tan, S. (2016). Managing rater effects through the use of FACETS analysis: The case of a university placement test. *Higher Education Research and Development*, *35*(2), 380–394. https://doi.org/10.1080/07294360.2015.1087381

Çok Yüzeyli Rasch Ölçümü (MFRM) Modeli ile Farklı Deneyim Puanları Arasında Derecelendirme Performansı

Giriş

Puanlayıcı aracılı değerlendirme, eğitim ortamında birçok yerde karşılaşılabilecek ve adayların karmaşık özelliklerini değerlendirmeye yönelik özellikle yüksek riskli değerlendirmelerde kullanılmaktadır. Bununla birlikte puanlayıcı kullanıldığı durumda puanlayıcıların yeterliği geçerliği doğrudan etkilemektedir. Puanlayıcılar değerlendirme prosedürlerini önemli bir şekilde takip etseler de puanlama performanslarında önyargılı davranabilirler. Ayrıca puanlayıcıların deneyimlerine bağlı olarak performasları da farklılık gösterebilmektedir. Bununla ilgili olarak ise alanyazında yapılmış çalışmalar bulunmaktadır (Ahmadi Shirazi 2019; Alp, Epner ve Pajupuu 2018; Attali 2016; Davis 2016). Bu çalışmalarda deneyimli puanlayıcıların puanlama sürecinde uzmanlıklarından daha fazla etkilendiği, acemi puanlayıcıların ise benzer kalitede puanlama yapamadıkları iddia edilmiştir. Sonuç olarak puanlayıcıların puanlama sürecinde birçok faktörden etkilendiği; bazılarının daha cömert bazılarının ise puanlamada daha katı davrandığı bilinmektedir. Bu puanlama süreçleri sonucunda ise sınava giren adayların puanları ciddi şekilde değişiklik göstermektedir. Özellikle sözlü sınavlarda puanlayıcıların değerlendirme prosedürlerine tam olarak uygun davranamadıkları, bu nedenle de adayların puanlarına puanlayıcısından kaynaklı hataların karışabileceği düşünülmektedir. Bu doğrultuda bu araştırma kapsamında sözlü bir sınavda puanlayıcıların puanlama performanslarının incelenmesi ve farklı puanlama deneyimlerine sahip değerlendiriciler ile deneyimsiz puanlayıcıların puanları arasında bir farklılık olup olmadığını belirlemek amaçlanmıştır. Araştırmanın temel problemleri bu doğrultuda şu şekildedir:

- Sözlü sınavların değerlendirilmesinde uygulanan puanlama ciddiyetinin belirlenmesi
- Deneyimli ve deneyimsiz puanlayıcıların performansları arasında bir fark olup olmadığının belirlenmesi

Yöntem

Nicel araştırma yönteminde yürütülen bu çalışmada ortaokul öğrencilerinin İngilizce sözlü sınavları puanlarının incelemesi gerçekleştirilmiştir. Toplam 30 ortaokulda görev yapan İngilizce öğretmeni puanlayıcı olarak çalışmaya dâhil edilmiştir. Öğretmenler yüksek riskli testleri puanlama konusundaki deneyimlerine dayanarak iki gruba ayrılmışardır. İlk grupta yer alan 15 öğretmen bu konuda deneyimsizken diğer gruptaki 15 öğretmen, yüksek riski testleri değerlendirme konusunda en az iki yıllık deneyime sahip kişilerdir. Araştırmanın verilerini öğrencilere uygulana sözlü test, sınava katılanların cevapları, puanlama anahtarı ve puanlama formu oluşturmaktadır. Sözlü testte genel görüşme, hikâye anlatımı ve tartışma olmak üzere üç görev bulunmaktadır. Sözlü anlatım testindeki görevleri cevaplandırmak üzere farklı yeterlik düzeylerine sahip 10 öğrenci seçilmiş ve öğrencilerin cevapları doğrultusunda simülasyon işlemi gerçekleştirilmitşir. Öğrencilerin üç görevdeki cevapları da kelime bilgisi, dil bilgisi ve iletişimsel yeterlik alanlarında değerlendirilmiştir. Puanlayıcıların tamamı, sınava katılan 10 öğrenciyi de puanlamışlardır. Verilerin analizi FACETS yazılımı kullanılarak gerçekleştirilmiştir (Linacre, 2014a). Veriler analiz edilmeden önce analiz için kullanılan Rasch modelinin varsayımları için MNSQ infit kullanılmış ve üç görev için de madde-uyum değerlerinin, standart hatanın ve PTMEA değerinin kabul edilebilir değerler arasında olduğu belirlenmiştir. Puanalam anahtarında kulanılan derecelendirme ölçeklerinin Linacre (2002) tarafından belirlenen altı temel koşulu karşıladığı ve tüm ölçeklerin çalışmada kullanmaya uygun olduğu tespit edilmiştir.

Sonuç ve Tartışma

Bu çalışmada öğrenciler, maddeler, puanlayıcılar ve alanlar olmak üzere dört facet bulunmaktadır. MFRM analizinde her birimi parametrelerine göre düzenlemek amacıyla Wirght haritası oluşturmaktadır. Homojenlik testi için ki-kare analizi, en az iki puanlayıcının puanlarının ciddiyetinin ki-kare değeri $\chi 2 = 266.2$, df = 29, p <.01 ile heterojen olduğunu göstermiştir. Puanlayıcı ayırma oranı 2,85'tir ve puanlayıcılar arasındaki ciddiyet farkının ölçüm hatasından neredeyse üç kat daha fazla olduğunu gösterir. Ayırma indeksi 4.13 olup istatistiksel olarak dörtten fazla puanlayıcı ciddeiyet katmanı olduğunu göstermektedir. Ayırma istatistiği güvenilirliği 0.89'dur, bu da puanlayıcı ciddiyetinin varyansının kesin olduğunu ve ölçüm hatalarından fazla etkilenmediğini göstermektedir. Bulgular, burada üç puanlayıcı ciddiyeti ve sadece bir ılımlı puanlayıcı olduğunu ortaya koymaktadır. Bu gözlem, puanlayıcıların puanlama anahtarını anlama konusundaki çeşitli yeteneklerinin ve konuşma becerilerini değerlendirme konusundaki aşinalıklarının bir sonucudur (Kang, Rubin ve Kermad, 2019). Uyumsuz sadece bir puanlayıcı vardır ve onun da puanları aşırı uyumsuz değildi. Araştırmanın ilginç olan bulgusu, uyumsuz puanlayıcının deneyimsiz gruptan olmasıdır. Bu sonuç, deneyimsiz puanlayıcılar arasında uyumsuz bir puanlamayı tespit eden Weilie (2018) ile benzerlik göstermektedir. Bununla birlikte, şaşırtıcı bir şekilde, Ahmadi Shirazi (2019), iki uyumsuz puanlayıcının beş yıldan fazla puanlama deneyimine sahip olan kişiler arasında olduğunu bulmuştur. Isaacs ve Thomson (2013), puanlama deneyimlerine dayanarak uyumsuz puanlayıcılar için net bir model olmadığını belirtmişlerdir. Kısacası elde edilen bu sonuçlar puanlayıcılar arasındaki uyumsuzluğun mutlaka puanlayıcıların deneyiminden kaynaklanmadığını göstermektedir. Bu çalışmadaki puanlayıcılar %64.5 düzeyinde uyum göstermişlerdir. Tüm puanlayıcıların, model tarafından beklenen puanlayıcılar arası uyumun kabul edilebilir sınırının üstünde puanlar verebildiklerini göstermektedir (Wu & Tan, 2016).

Araştırmada daha sonra puanlayıcıların ciddiyeti, bağımsız örneklem t-testi ile her iki grup için karşılaştırılmıştır. Tablo 1'deki bulgular, iki grup arasında istatistiksel olarak anlamlı bir fark olduğunu göstermiş (t-değeri = -0.96, df = 28, p <0.01) ve deneyimli grup ile deneyimsiz grup arasında fark olmadığını belirten yokluk hipotezinin reddedilmesini sağlamıştır. Bu sonuç Attali (2016) bulgularıyla tutarlıdır; Davis (2016) ve Huang ve diğ. (2018), farklı puanlama deneyimine sahip değerlendiricilerin, farklı çalışmalar bağlamında heterojen puanlama yaptıklarını ortaya koymuştur. Sonuçlar arasındaki bu tutarlılık, puanlayıcıların puanlama deneyimlerinin nasıl tanımlandığında bağlı olarak da değişebilir. Biz bu çalışmamızda deneyimli puanlayıcı olarak yüksek riskli testlerde puanlama deneyimine sahip olan kişileri tanımladık.

Tablo 1. Deneyime Göre Puanlayıcı Ciddiyetinin Farklılıkları

Gruplar	N	Ortalama loglar	Standart Sapma	Standart Hata	Değerlendiriciler arası anlaşma
Deneyimsiz değerlendiriciler	15	-0.14	1.00	0.26	58.9%
Deneyimli değerlendiriciler	15	0.14	0.47	0.12	70.0%

t değeri = -0.96; *df* = 28; *p*<0.01

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2020; 11(2); 163-179

Adaptation of the Self-efficacy Beliefs in STEM Education Scale and Testing Measurement Invariance across Groups *

Cansu DEMİRBAĞ ** Serkan ARIKAN *** Ebru Zeynep MUĞALOĞLU ****

Abstract
Academic performance on science, technology, engineering, and mathematics (STEM) education is important for the economic development of countries. From the perspectives of social cognitive theory, one of the predictors of academic performance is self-efficacy. In order to measure middle school students' self-efficacy beliefs in STEM education, STEM Competency Beliefs scale was developed in English originally by Chen, Cannady, Schunn, and Dorph (2017). In this study, it is aimed to adapt the English scale into Turkish and to provide evidence regarding reliability and validity. Throughout the adaptation process, forward and backward translation was completed. In the pilot study (n = 77), the reliability of the data and the clarity of the statements in the Turkish version of the scale was examined. In the main study, the Turkish version was administered to 330 middle school students to investigate the psychometric properties of the scale. The results pointed out that the scores obtained by the Turkish version of the scale had good internal consistency. Regarding the dimensionality of the scale, in contrast to the original version, the adapted scale showed a two-dimensional structure. Measurement invariance findings for gender groups supported configural and metric invariance, whereas scalar invariance was partially achieved. Measurement invariance findings for career choice groups supported configural, metric, and scalar invariance. Scale scores of students were estimated using multidimensional Item Response Theory. The findings suggested that the scale can be utilized for STEM-related research to assess the competency beliefs of students.

Key Words: Self-efficacy beliefs, scale adaptation, confirmatory factor analysis, measurement invariance, multidimensional item response theory.

INTRODUCTION

Science, technology, engineering, and mathematics (STEM) education is the integration of these disciplines (Breiner, Harkness, Johnson, & Koehler, 2012; Tsupros, Kohler, & Hallinen, 2009) in order to deal with real-world problems (Johnson, Peters-Burton, & Moore, 2016; National Research Council-NRC, 2014). STEM education is substantial for countries in terms of three interconnected aspects: competitiveness in the global market, needs for innovation, and jobs of the future (Atkinson & Mayo, 2010; English, 2016; Johnson et al., 2016). One of the ways to stay competitive in global markets for countries is maintaining development in STEM disciplines. Science- and technology-based innovation enforces countries in the global market by increasing exports (Atkinson & Mayo, 2010). This kind of innovation is only possible with a workforce educated in science, technology, engineering, and mathematics content (Atkinson & Mayo, 2010). It is predicted that in the future one out of three jobs will be STEM-integrated or strongly related to STEM fields. Hence, students need to be educated with integrated STEM approach as candidates for the future workforce (English, 2016).

Similarly, Turkey, as a developing country, emphasizes the importance of STEM education for its' economic growth (TÜSİAD, 2019). Turkey needs a qualified and talented workforce educated through

* This study is a part of Master Thesis entitled "A Turkish Adaptation of The Stem Competency Beliefs Instrument" completed within Boğaziçi University Institute for Graduate Studies in Social Sciences.

** Graduate Student, Boğaziçi University, Faculty of Education, İstanbul-Turkey, cansu.demirtas@boun.edu.tr, ORCID ID: 0000-0001-9072-0375
*** Assist. Prof., Boğaziçi University, Faculty of Education, İstanbul-Turkey, serkan.arikan1@boun.edu.tr, ORCID ID: 0000-0001-9610-5496
**** Assoc. Prof., Boğaziçi University, Faculty of Education, İstanbul-Turkey, akturkeb@boun.edu.tr, ORCID ID: 0000-0001-7766-4743

To cite this article:
Demirbağ, C., Arıkan, S., & Muğaloğlu, E. Z. (2020). Adaptation of the self-efficacy beliefs in STEM education scale and testing measurement invariance across groups. *Journal of Measurement and Evaluation in Education and Psychology, 11*(2), 163-179. doi: 10.21031/epod.675240

Received: 15.01.2020
Accepted: 26.04.2020

STEM fields to achieve the goals of 2023. Preliminary actions have been done, such as changing the national curriculum (Ministry of National Education-MEB, 2018a) and opening STEM institutions and centers to empower STEM education (Colakoğlu & Gökben, 2017). Moreover, research about STEM studies and developing STEM-related master and doctorate programs have been increasing (Akgündüz et al., 2015).

Self-efficacy beliefs are regarded as one of the variables that play a key role in academic achievement (Jinks & Lorsbach, 2003; Kanny, Sax, & Riggers-Piehl, 2014; Nelson & Ketelhut, 2008) and career persistence (Green & Sanderson, 2018) in STEM fields. It is significant to improve self-efficacy and academic achievement of students in STEM fields to fulfill the STEM-related jobs. Even though the number of STEM education research has gained acceleration both at international level (Atkinson & Mayo, 2010; Breiner et al., 2012; English, 2016; Johnson et al., 2016; Tsupros et al., 2009) and in Turkey (Han, Capraro, & Capraro, 2016; Hacıoğlu, Yamak, & Kavak, 2016; Yerdelen, Kahraman, & Taş, 2016), to the best of our knowledge, there is not a valid scale to assess the STEM self-efficacy beliefs in Turkey.

Firstly, the present study aimed at adapting the English version of the STEM Competency Beliefs scale into Turkish and validating the adapted version. Secondly, the study compared the participants' self-efficacy beliefs on STEM education in terms of their gender, school type, and career choices in a Turkish context. Finding significant differences between school types (private vs. public) and career choices (stem related and not-stem related) could be considered as additional validity evidence (Sireci & Sukin, 2013) as these groups are expected to be different in their competency scores due to the resources and student motivation, respectively.

Having a valid scale to assess STEM self-efficacy beliefs in Turkish is significant for researchers and educators to investigate individual's self-efficacy on STEM and its relationships with other crucial variables such as academic performance in STEM or interest towards STEM fields in Turkey. Moreover, having a STEM Competency Belief scale in Turkish enables researchers, teachers and policymakers to evaluate STEM programs and identify the learner characteristics in terms of STEM self-efficacy in Turkey. Comparing STEM competency beliefs of gender groups in Turkey is also expected to extend the literature.

Self-efficacy Beliefs in STEM Education

Self-efficacy is defined as the capability of an individual's point of view for himself/herself to perform at a level of proficiency (Bandura, 1999) and interchangeably used perceived self-competence (Zimmerman, 1995). Self-efficient people are more resilient, solution-oriented, hard workers (Pajares & Miller, 1997), active in the control of time, better at task focus (Bouffard-Bouchard, Parent, & Larivee, 1991), self-regulated, more efficient in the use of problem-solving strategies and in the management of working time (Zimmerman, 2000). Bandura (1999) also explained that self-efficient people perceived failure differently than less self-efficient people. They regard failure to insufficient effort, weak strategies, or conditions. These features of self-efficient people play a key role in their performance (Bandura, 1999; Bouffard-Bouchard et al., 1991).

Beliefs about self-efficacy influence how much students learn (Vincent-Ruz & Schunn, 2017). For instance, Nelson and Ketelhut (2008) investigated ninety-six middle school students' self-efficacy and their performance in learning science in a virtual environment. As a result of the study, it was indicated that students with lower levels of self-efficacy did not perform as well as students with higher levels of self-efficacy. Bandura (1997) emphasized that the relationship between self-efficacy and performance is reciprocal. In other words, if people are self-efficient, their characteristics help them to be successful in related tasks. Achieving tasks boosts their self-efficacy, which leads to working harder and targeting more difficult tasks. Working harder helps to achieve new tasks that continue with better performance and higher self-efficacy. Moreover, Hidi and Ainley (2008) emphasized a positive relationship between interest and self-efficacy. The more students believe themselves, the more they are interested in their subjects. Thus, educators are required to help learners to experience better feelings and improve their beliefs about themselves. It helps students continue to work on or reengage

with activities, ideas, objects and so on, and to increase knowledge and a stored value (Hidi & Ainley, 2008).

Beliefs about capabilities function as an important role that influences science or non-science related majors and career choices (Hackett & Betz, 1982). Durik, Vida, and Eccles (2006) examined how the 10th graders' self-concept of ability on English/reading was related to their career choices. The results showed that the subject-oriented self-concept of ability predicted future career preferences of 10th graders. Gainor (2006) also emphasized that people choose careers in areas where they believe that they are good at doing it well.

Studies found that females have lower self-efficacy towards STEM fields (Tellhed, Backström, & Björklund, 2017). Females do not believe that they can accomplish STEM fields because of the lack of role models and social or verbal persuasions (Zeldin, Britner, & Pajares, 2008). Self-doubts, lower performance expectations, male-dominated fields, social persuasions and vicarious experiences about STEM fields, individual backgrounds, family influences and expectations, perceptions towards STEM fields, psychological values, factors, and preferences are related with females' lower interests towards STEM fields (Kanny et al., 2014; Tellhed et al., 2017; Zeldin et al., 2008). Lower self-efficacy beliefs of females towards STEM is needed to overcome to reduce gender segregation in the field. One of the ways for increasing females in the area is increasing their self-efficacy for STEM careers (Tellhed et al., 2017).

Self-efficacy is a personal state which can change especially based on positive personal outcomes. As Jenson, Petri, Day, Truman, and Duffy (2011) stated STEM self-efficacy is an important focus and worthy of observation. Therefore, to assess STEM self-efficacy, many scales have been developed over the years (e.g., Dawes, Horan, & Hackett, 2000; Lent, Brown, & Larkin, 1986). In 2014, Milner, Horan, and Tracey (2014) argued that most of the scales have validity issues, and they developed the STEM Career Self-Efficacy Test. Pieces of evidence were presented to claim that the scale can be accepted as a valid instrument to measure self-efficacy in engaging STEM activities (Milner et. al., 2014). However, the scale is not applicable to middle school students who are expected to learn STEM fields at schools. In 2017, the STEM Competency Beliefs scale was developed for middle school students in Activation Lab in the USA (Chen, Cannady, Schunn, & Dorph 2017). Activation Lab gathers academicians from various universities of the USA. They aim to increase young people's understanding and appreciation of STEM to prepare them for future challenges. One of the main research areas in Activation Lab is to develop scales to measure significant variables for STEM education, such as Science Competency Scale (Chung, Cannady, Schunn, Dorph, & Vincent-Ruz, (2016) and STEM Competency Belief scale (Chen et al., 2017). The STEM Competency Belief scale was developed to assess an individual's STEM Competency Beliefs. Cannady stated that the scale was also adapted into different languages like Spanish and African (M. Cannady, personal communication, November 12, 2018). As the original scale was developed very recently, there is not any publication yet based on this scale. Moreover, Smith (2019) adapted the original scale to measure technology competency beliefs. She applied the adapted version to investigate the effect of a coding instruction to seventh graders' self-efficacy in technology.

Present Study

In a decade when STEM has gained popularity and been studied from different perspectives, it is crucial to assess the self-efficacy of students for STEM fields. One of the scales to assess middle school students' self-efficacy in STEM education is the STEM Competency Beliefs scale. The scale was developed by Chen et al. (2017) in English. The purpose of the present study was twofold. First, to adapt the scale into Turkish and to test the factor structure of the STEM Competency Beliefs scale with the Turkish sample. The second purpose was to test whether the factor structure of the scale had measurement invariance across gender groups and career choice groups in the Turkish sample. The research questions of this study are:

1) Does the factor structure of the adapted STEM Competency Beliefs scale similar to the original scale?
2) Are the configural, metric, and scalar parameters invariant across girls and boys?
3) Are the configural, metric, and scalar parameters invariant across students who want to follow stem-related and not stem-related careers?
4) Is there any significant difference between students' scale scores on gender groups, career groups, and school types?

METHOD

This study primarily aimed to adapt STEM Competency Beliefs scale into Turkish and to test measurement invariance for the factor structure of the STEM Competency Beliefs scale. Therefore, the adaptation part could be named as a descriptive study and measurement invariance part could be named as a correlational study. Detailed information about participants, data collection instrument and data analysis are presented below.

Participants

For the pilot and the main study, two different sample groups were used. All the students were science center visitors taken by their schools as a school trip to attend workshops; therefore, the sampling method was the convenience sampling. These workshops were held in a science center in İstanbul which belongs to a Municipality. Seventy-seven students (4^{th} to 8^{th} graders) participated in the pilot study. The participants consisted of 32 male (42%) and 45 female (58%) students. Seven of the participants (9%) were from private schools, and 70 of them (91%) were from public schools.

Participants of the main study were 330 students coming from different schools as visitors to the science center. Among these 330 students, 4 of them did not provide all responses to the items. Therefore, after listwise deletion, all the analyses were conducted based on 326 students (2 females and 2 males; 3 public and 1 private school). The gender percentages of the students were regarded as balanced, consisting of 157 females (48%) and 169 males (52%). Also, students who participated in the study were coming from different school types as public schools (n = 302, 93%) and private schools (n = 24, 7%). The majority of the students were 7th graders. Among these students, 161 of them (49%) stated that they want to have STEM-related careers, whereas 165 of them (51%) do not want to follow STEM-related careers. According to student ratios of gender groups, school types, and students' choices of future careers, and the way these students were brought to the center, the sample could be considered as not biased.

Data Collection Instrument

The STEM Competency Belief scale is a 12-item 4-point Likert-type scale (Chen et al., 2017). The survey was designed for 10-14-year-old respondents to assess an individual's STEM Competency Beliefs. The reliability of the STEM Competency Beliefs Scale was good (Cronbach's Alpha = .83; polychoric Alpha = .87) based on a data collected from a sample of 205 middle school youth (Chen et al., 2017). Two of the items were listed below as sample items:

"I can do math problems I get in the class."

"I am the technology expert in the house."

Data Analysis

The scale adaptation process included the following stages: scale adaptation, piloting, reliability and validity analysis, and testing measurement invariance for gender groups and career choice groups.

Scale adaptation

Methodology in translation and adaptation of a scale has enhanced rapidly in last 25 years. The reasons behind this rapid development are based on four issues including interest in cross-cultural psychology (van de Vijver & Hambleton, 1996), international comparative studies in education, worldwide exams, and fairness in testing for language preferences (Hambleton, Merenda & Spielberger, 2012; International Test Commission-ITC, 2017;).

Translation and adaptation are two major terms used in the field. Compared to the test translation, the test adaptation is a more preferred, more reflective, broader, and commonly used term (Hambleton et al., 2012; ITC, 2017). During the application of test adaptation, a variety of activities are required, such as deciding whether the same construct occurs in different languages, determining translators, deciding accommodations, adapting the tests, and checking for equivalence. On the other hand, the test translation is only one of the steps that happen in the adaptation. This step is language translation from one to another. However, a test adaptation requires thinking deeply in terms of cultural, psychological and linguistic issues (Hambleton et al., 2012). Briefly, translation and adaptation have different meanings, and the adaptation is a more comprehensive term.

ITC (2017) guideline grouped the steps of the test adaptation process as before, in progress, and after. According to the guideline, before the adaptation, three steps are suggested for experts: obtaining permission from test developers, evaluating the similarities between cultures, and minimizing the cultural and linguistic differences. In the progress part of the adaptation, five steps are emphasized: ensuring the minimal cultural differences, using appropriate design methods to maximize suitability, providing evidence that the test is the same for intended populations, providing evidence for the structure of the test, collecting data to complete necessary revisions. In the last part, four steps are needed to be completed after the adaptation process: determining the sufficient size of the sample, providing statistical evidence for construct equivalence, providing evidence for reliability and validity analysis, and using appropriate data analysis procedure. In addition to the steps mentioned here, scoring and documentation are emphasized in the guideline (ITC, 2017).

For the adaptation process, two main design methods appear in the literature, namely forward and backward translation. The forward translation is a process that one or more translators adapt the test from the source language to the target language. Backward translation has three main processes in itself. Firstly, a test is translated from the source language to target language by determined translators. Then, different translators translate the test from target language back to the source language. Finally, these two forms of the test as source language and back-translated version are compared for equivalence (Hambleton et al., 2012). The backward translation allows the researcher to compare two forms in a more objective level.

For the adaptation of the STEM Competency Beliefs scale, preconditions were completed before the study. Firstly, permission was granted for the adaptation of the STEM Competency Beliefs scale into the Turkish (M. Cannady, personal communication, November 12, 2018). Then, cultural similarities and differences were evaluated by the research team, including an associate professor in science education, an assistant professor in assessment and evaluation, and the researcher. Finally, forward translation, backward translation, and final version editing were performed.

Forward translation: For the forward translation, the scale was translated from English to Turkish. Translators were 5 years experienced English teacher and 7 years experienced English interpreter. Each translator worked independently, and translated forms were collected in an excel document. The research team compared the translations, discussed STEM-related terms, and the scale was formed in Turkish. For example, the research team discussed "After school science club" and decided to translate as "science and technology club" which is a term in the National Education Social Activities Program Students' Club (MEB, 2009).

Backward translation: To achieve backward translation, two additional translators translated the scale from Turkish to English. These translators were a Turkish scholar who lived in England for 25 years and an American author who has been living in Istanbul for 14 years. Back-translated forms were

again collected in an excel document, and the research team investigated the similarities between the original form of the scale with back-translated form. After all, the research team reached a consensus for the back-translated scale.

Final version editing: As a final step, a linguist expert who is a doctorate student in a Learning Science program and a Turkish language editor compared the back-translated version of the scale and the original one. After some smooth changes on the adapted scale, the adapted Turkish version was finalized.

Piloting the adapted version of the scale

A pilot study was conducted to check the clarity of the items from students' perspectives. There were 2 additional questions at the end of the survey: "Is there any question that you struggle to understand?" and "if yes, which question(s) were they?" to identify problematic statements. Additionally, Cronbach's Alpha value and corrected item-total correlations were estimated to flag problematic items. Related revisions were made as a result of the pilot analysis.

Reliability analysis of final data

The reliability of the scale was tested using Cronbach's Alpha internal consistency coefficient. Cronbach's Alpha value above .70 is acceptable, above .80 is good, and .90 and above is excellent. Results that are closer to 1 mean higher internal consistency (George & Mallery, 2001). In the item level, the corrected-item total correlations were reported. Items with low correlations (less than .30) are considered as problematic items (Field, 2013), and these items are investigated to detect the source of the problem.

Validity analysis of the final data

For the validity analysis, confirmatory factor analysis (CFA) was conducted. CFA is one of the forms of factor analysis to test whether the hypothesized structure fits the collected data well or not (Urdan, 2010). In order to evaluate the goodness of the fit of the data for the proposed model, fit indices are used. CFI (Comparative Fit Index), TLI (Tucker Lewis index) and RMSEA (Root Mean Square Error of Approximation) are widely used fit indices that are less sensitive to the sample size. CFI and TLI values over .95 and RMSEA value smaller than .06 is accepted as a good fit (Ullman, 2001). CFA analysis for the study was conducted with MPLUS 7.4 (Muthén & Muthén, 2015) using the Weighted Least Square estimation method. One dimensional structure proposed in the English version was tested with the data collected by the adapted Turkish version. Multivariate normality, outliers, and sample size assumptions were checked to conduct CFA (Ullman, 2001).

When the student data does not fit the hypothesized structure, exploratory factor analysis (EFA) could be used to investigate the communalities among items. EFA using principal axis factor extraction technique with direct oblimin rotation was conducted as items could be correlated with each other. An item that has 0.400 or less item loading to its primary factor is considered as a problematic item. Also, if an item is loaded to at least two factors at the same time (factor loading difference of an item to a primary factor and other factor is less than .10), that item is also called problematic item (Field, 2013).

Item response theory scaling

Item response theory (IRT) scaling was conducted to estimate students' ability on the latent variables. Generally, IRT requires the data to be unidimensional (Hambleton & Jones, 1993). In the case of violating unidimensionality, multidimensional IRT estimations are available (Reckase, 2009). IRTPRO 4.2 (Cai, Thissen & du Toit, 2017) software was used to estimate the student ability as the

software is capable of conducting unidimensional and multidimensional IRT. Bock-Aitkin Expectation-Maximization estimation method was used.

Measurement invariance of final data

Measurement invariance analysis for gender groups and career choice groups were conducted to test whether the same construct was being measured across groups. As the number of students from private schools was not enough to estimate the parameters, measurement invariance analysis for school type was not performed. Having measurement invariance across gender or career choice groups implies that the scale scores of boys and girls, or students who want stem-related and not stem-related careers are comparable. The measurement invariance is tested comparing fit results of nested models: configural, metric, and scalar models. In the configural model, whether the same factor structure exists across groups is tested. In this model, factor loadings and thresholds are freed to be different across groups. In the metric model, factor loadings were constrained to be equal across groups, but the thresholds could take different values. In the scalar model, both factor loadings and item thresholds are constrained to be equal for groups (Milfont & Fischer, 2010; Vandenberg & Lance, 2000). Measurement invariance is assessed by comparing ΔCFI and ΔRMSEA values with cutoff criteria ($\Delta CFI \leq .01$, $\Delta RMSEA \leq .015$) suggested by Chen (2007), and Cheung and Rensvold (2002).

RESULTS

Pilot Study of the Scale

In the pilot study, items were administered to 77 students to test the clarity and fluency of the statements mainly. There were 2 additional questions at the end of the survey: "Is there any question that you struggle to understand?" and "if yes, which question(s) were they?" Seventy-two students stated that they could understand the statements clearly, and five students indicated that they had a problem to understand some items. These answers were used to determine if the statements need any changes or improvements before finalizing the Turkish version. For instance, one child expressed that item 2 was difficult for her/him because the word *website* was not familiar to him. Then, the word *website* changed as *internet sitesi* for the main study. Cronbach's Alpha coefficient of the data was found as .75. Corrected item-total correlations were between .28 (item4) to .60 (item12) which were acceptable values.

Reliability Analysis of the Final Scale

The reliability analysis of the final form of the 12-item scale pointed out that Cronbach's Alpha coefficient was .83, which implied the data had good internal consistency. Table 1 showed that the corrected-item total correlation of each item was higher than .30, which means that there were no problematic items in terms of item discrimination.

Table 1. Corrected Item-Total Correlations of Final Study

Item	Corrected Item-Total Correlation	Cronbach's Alpha If Item Deleted
Item 1	.51	.81
Item 2	.49	.82
Item 3	.49	.82
Item 4	.37	.83
Item 5	.50	.82
Item 6	.43	.82
Item 7	.52	.81
Item 8	.50	.82
Item 9	.57	.81
Item 10	.49	.82
Item 11	.52	.81
Item 12	.48	.82

Confirmatory Factor Analysis

The original scale was shown to have a one-factor structure by the scale developers. Therefore, in the CFA, the adapted version of the scale was hypothesized to have a one-factor structure. The assumptions of multivariate normality were tested by drawing a histogram and estimating skewness and kurtosis. As histogram, and skewness (-.28) and kurtosis (-.30) values implied, the data were distributed normally. There was no outlier in the data. The ratio of sample size to the number of the variable was 27.5, which implied that the sample size was sufficient. The ratio of 1 to 10 is considered as enough sample size (Bentler & Chou, 1987). The fit statistics obtained through CFA was not acceptable for the one-factor model as shown in Table 2 (CFI = .890 < .950; TLI = .866 < .950; RMSEA = .117 > .060).

Table 2. One-Factor Confirmatory Factor Analysis

	χ^2	df	χ^2 / df	CFI	TLI	RMSEA
Model 1	295.946	54	5.480	0.890	0.866	0.117

Hence, exploratory factor analysis (EFA) was conducted to understand the structure of the Turkish version. Principal axis factoring (PAF) with oblimin rotation was performed for the EFA. Kaiser-Meyer-Olkin measure of sampling adequacy value of .863 indicated that the proportion of variance in the items might be caused by the underlying factor. Bartlett's test of sphericity (p < .05) showed that the correlation matrix was different from an identity matrix. Therefore, the data was appropriate for conducting the exploratory factor analysis. As shown in Table 3, the data had a two-factor structure where items 1, 8, and 9 were loaded to a different factor.

The items that were loaded to a new factor were listed below. These three items include statements regarding mathematics, whereas the other nine items focus on science, technology, and engineering. Hence, the primary factor was called self-efficacy related to science-technology-engineering (STE), and the second factor was called self-efficacy for mathematics (Math). Items loaded to the second factor are listed below.

Item 1: "I can do math problems I get in class."

Item 8: "I think I am very good at Explaining my solutions to math problems."

Item 9: "I think I am very good at: Solving problems"

Table 3. Exploratory Factor Analysis

Item	Factor	
	1	2
Item 10	**.649**	.066
Item 11	**.633**	.022
Item 5	**.585**	.008
Item 6	**.564**	.067
Item 12	**.508**	-.060
Item 4	**.454**	.032
Item 3	**.434**	-.163
Item 2	**.428**	-.153
Item 7	**.416**	-.234
Item 8	-.034	**-.785**
Item 1	-.016	**-.776**
Item 9	.143	**-.653**

As the data structure in PAF suggested a two-factor structure, a CFA with two factors was reconducted. The two-factor model improved the fit statistics impressively as shown in Table 4 (CFI = .974 > .950; TLI = .968 > .950; RMSEA = .057 < .060). This finding showed that the STEM Competency Beliefs

scale had the two-factor structure for the Turkish data as science-technology-engineering is the first factor, and mathematics is the second factor.

Table 4. Two-Factor Analysis

	χ^2	df	χ^2 / df	p	CFI	TLI	RMSEA
Model 2	109.466	53	2.065	0.000	.974	.968	.057

Measurement Invariance

Configural, metric and scalar invariance of the scale across gender groups and career choice groups were evaluated (See Table 5 and 6). For school type, as there were a limited number of students in one group (24 students in from private school), measurement invariance analysis could not be achieved. Configural invariance results across gender groups indicated that the fit indices were good (TLI = .971, CFI = .975, RMSEA = .058). This means that the factor structure of the scale was similar for boys and girls. Metric invariance analysis showed that the change in the fit statistics supported the invariance (ΔCFI = .001, ΔRMSEA = -.003). Having metric invariance means that in addition to the factor structure, the factor loadings were equivalent across gender groups. Scalar invariance results showed that the change in the CFI was higher than allowed, whereas, for RMSEA, the change was within an acceptable range (ΔCFI = -.016, ΔRMSEA = .006). Modification indices suggested that this problem could be due to item 7. Freeing thresholds of item 7 for boys and girls resulted in better and accepted change in fit statistics (ΔCFI = -.010, ΔRMSEA = .002). This finding means that except item 7, item thresholds were invariant, and mean scores of males and females were comparable. Item 7 is "I think I am very good at: Giving evidence when I tell my opinion." Therefore, partial scalar invariance was supported for gender groups.

Configural invariance results across career choice groups indicated that fit indices were good (TLI = .961, CFI = .969, RMSEA = .063). This means that the factor structure of the scale was similar for students who want to follow STEM-related or not STEM-related careers. Metric invariance analysis showed that the change in the fit statistics supported the invariance (ΔCFI = .002, ΔRMSEA = .005). Having metric invariance means that besides the factor structure, the factor loadings were equivalent across career choice groups. Scalar invariance results showed that the changes in the CFI and RMSEA were also within acceptable ranges (ΔCFI = .000, ΔRMSEA = .009). This finding suggested that the mean scores of career choice groups are comparable.

Table 5. Measurement Invariance Analysis of the Scale for Gender Groups

	χ^2	*df*	χ^2 / *df*	TLI	CFI	RMSEA	ΔCFI	ΔRMSEA
Configural	164.13	106	1.55	.967	.974	.058 (.040; .075)	-	-
Metric	172.32	116	1.49	.971	.975	.055 (.036; .074)	.001	-.003
Scalar	230.88	138	1.67	.960	.958	.064 (.049; .079)	-.016	.006
Scalar new	215.41	135	1.60	.965	.964	.060 (.045; .075)	-.010	.002

Note. χ^2 = Chi-square, df = degrees of freedom, TLI = Tucker Lewis index, CFI = comparative fit index, RMSEA = root mean square error of approximation; CI = confidence interval, ΔCFI = change in values of CFI, ΔRMSEA = change in values of RMSEA. Scalar new: Thresholds of items 7 is freed.

Table 6. Measurement Invariance Analysis of the Scale for Career Choices

	χ^2	df	χ^2 / df	TLI	CFI	RMSEA	ΔCFI	ΔRMSEA
Configural	173.68	106	1.64	.961	.969	.063 (.045; .079)	-	-
Metric	178.68	116	1.54	.967	.971	.058 (.040; .074)	.002	.005
Scalar	203.80	138	1.48	.971	.969	.054 (.037; .069)	.000	.009

Note. χ^2 = Chi-square, df = degrees of freedom, TLI = Tucker Lewis index, CFI = comparative fit index, RMSEA = root mean square error of approximation; CI = confidence interval, ΔCFI = change in values of CFI, ΔRMSEA = change in values of RMSEA.

Comparative Analyses

Comparative analyses were conducted to test mean score differences of related groups (gender, school type, and career choices). The scores used in these comparisons were estimated using multidimensional IRT scaling. As all subgroup scores were normally distributed, a parametric test of group comparison was chosen. For the first comparison, Science, Technology, and Engineering (STE) and Mathematics (Math) score means were compared for gender groups, excluding item 7. Table 7 shows the mean score of boys and girls for STE and Math factors. Independent sample t-test showed that the mean score difference of self-efficacy on Math for boys and girls was not statistically significant ($p > .05$; d = 0.12). A similar result was found for STE mean scores of boys and girls ($p > .05$; d = 0.21).

Table 7. Mean Scores of Gender Groups

	Gender						95% CI for Mean Difference			
	Male			Female						
	M	SD	N	M	SD	n		t	df	Cohen's d
STE	.09	.90	169	-.10	.89	157	-.38; .01	1.88	324	.12
Math	.05	.96	169	-.06	.83	157	-.31; .08	1.13	324	.21

For the second comparison, STE and math factor score means were compared for public and private schools. The mean score differences between public and private school students were statistically significant for both STE and Math, as showed in Table 8. Levene's test for equality of variances indicated that the variances were equal (p =.35 for STE and p =.07 for Math). In order to assess the magnitude of the differences, effect sizes were calculated (d = 0.83 for STE, and d = 1.27 for Math). The differences between public and private school groups were significant, with large effect sizes for both STE and math (Cohen, 1988).

Table 8. Mean Differences in School Type

	School Type						95% CI for Mean Difference			
	Public			Private						
	M	SD	N	M	SD	n		t	df	Cohen's d
STE	-.05	.90	302	.64	.75	24	-1.07; -.32	-3.68***	324	.83
Math	-.07	.89	302	.88	.57	24	-1.31; -.59	-5.18***	324	1.27

*$p < .05$, **$p < .01$, ***$p < .001$

As the third comparison, the mean scores of students according to their career choices (STEM-related vs. not STEM-related) were compared. Table 9 demonstrates that there are statistically significant differences between the groups. Cohen's d was calculated for the group and obtained 0.38 for STE and 0.41 for Math. It shows the group mean scores are not equal, and they have a medium effect size.

Table 9. Mean Differences on Career Choices

	Career Choices						95% CI for Mean Difference			
	STEM Related			Not-STEM Related						
	M	SD	N	M	SD	n		t	df	Cohen's d
STE	.17	.93	161	-.17	.85	165	-.54; -.15	-3.46**	324	.38
Math	.18	.87	161	-.18	.90	165	-.55; -.16	-3.64***	324	.41

*$p < .05$, **$p < .01$, ***$p < .001$

DISCUSSION and CONCLUSION

This study contributes to the literature by adapting the STEM Competency Beliefs scale to the Turkish. Providing evidence regarding reliability and the validity of the adapted STEM Competency Beliefs

scale is expected to enable scholars to use the scale in the Turkish context. Providing measurement invariance results before comparing mean scores of scales for subgroups is also important to exemplify the procedure in comparative studies. In this respect, this study fills a gap by providing an adapted version of the newly emerging Stem Competency Beliefs scale.

An important difference between the English original and Turkish adapted scale emerged in the dimensionality of the scale. While the original scale was reported to have a one-factor structure, the Turkish scale was shown to have a two-factor structure. Item 1, 8, and 9 were loaded to a different factor, which was closely related to Math-related self-efficacy. The rest of the items were related to science, technology, and engineering. Cannady stated that the scale was also adapted into different languages as Spanish and African (M. Cannady, personal communication, November 12, 2018), and those data also showed a unidimensional structure. It can be argued that there is a sharp distinction in STEM perceptions of Turkish students as considering math in one group, and science, technology, and engineering projects in the other group. This distinction is not an expected interdisciplinary view proposed by the STEM theory. The reason for this distinction could be that Turkey does not have a direct STEM action plan, whereas many countries have a concrete strategy plan and action (MEB, 2016). Hence, students in Turkey have difficulty in perceiving STEM as a whole. Besides that, in the latest revisions of the curriculum in Turkey, there is a statement emphasizing the "science, technology, engineering" in one hand, and mathematics on the other hand (MEB, 2018a, 2018b). This might be one of the plausible explanations of why students consider STEM fields in two distinct groups. Also, studies in Turkey supported the idea that STEM is not taught in an integrative way in the schools (Baran Canbazoglu-Bilici, Mesutoglu, & Ocak, 2016; Colakoglu, 2016; Ercan, Altan, Taştan, & Dağ, 2016; Han, Yalvac, Capraro, & Capraro, 2015). All the issues mentioned here may lead students not to comprehend STEM in the actual manner.

As the mean scores of boys and girls are compared frequently throughout the scales, providing evidence regarding measurement invariance is important to get valid inferences. The measurement invariance findings showed that configural and metric invariance was supported whereas scalar invariance could be achieved freeing item 7 across gender groups. This means that the factor structure of the scale and the factor loadings were similar for boys and girls. Except for item7, threshold values to endorse statements were also similar. Therefore, excluding item 7, mean scores of boys and girls on these factors are comparable. Item 7 is related to giving evidence about opinions. This finding implies that for boys and girls, providing evidence for their opinions could have a different meaning. Similarly, measurement invariance results for student groups according to their career choices (STEM-related vs. not STEM-related) suggested that the mean scores of career choice groups could be comparable.

Comparative analysis results showed that the mean score difference of self-efficacy on Math for boys and girls was not statistically significant, as well as STE mean scores. The effect sizes also supported these findings. On the contrary to the literature (Hackett & Betz, 1982; Tellhed et al., 2017; Zeldin et al., 2008), no major differences were observed between mean scores of both STE and Math factors in Turkey. The studies in the literature generally were related to high school or older students. Hence the lower ages of the participants of this study might be an explanation for a different pattern of the findings in Turkey. It can be stated that female students are as comfortable as male students towards STEM fields in Turkey.

Secondly, it was found that students at private schools had higher self-efficacy towards STEM compared to students at public schools. This finding might be related to learning opportunities, teachers' professional development, and class size differences between school types. Many private schools promote STEM education, have STEM laboratories, and invest in robotics and technology competitions at the national and international levels. These activities and opportunities may have a positive influence on private school students. This finding is also consistent with the literature (Chittum, Jones, Akalin, & Schram, 2017; John, Bettye, Ezra, & Robert, 2016; Monterastelli, Bayles, & Ross, 2008). Additionally, teacher-related variables are an important predictor for students' academic performance (Corlu, Capraro, & Capraro, 2014). Teachers working in private schools have more opportunities to take STEM-related professional in-service training. On the other hand, public

school students mostly depend on the individual efforts of their teachers. Lastly, class size might be an explanation for the differences because private schools have smaller class sizes than public schools. Other significant differences in the scale scores were found between students who want a STEM-related career and who do not want a STEM-related career. It was observed that students who want to follow STEM-related careers had higher self-efficacy beliefs on STEM. Having an interest in STEM fields as a future career might affect these students' self-efficacy in STEM fields.

Finding significant differences between private and public school students' mean scores and between mean scores of students who want a STEM-related career or not strengthen the validity of the scale. This scale could differentiate scale scores of students who have better opportunities in private schools and who have limited resources in public schools in terms of STEM education. Additionally, this scale could assign different scores for students who want to pursue a career in STEM-related fields and for students who are not willing to pursue such a career. These findings are additional evidence for the validity of the scale (Sireci & Sukin, 2013). Therefore, this reliable and valid scale is expected to contribute to the STEM self-efficacy research in the Turkish context.

Limitations

The main limitation of the study was related to the sampling procedure. As convenience sampling was used, the generalizability of the findings could be limited. Testing the structure of the scale with another sample would provide additional evidence regarding the structure.

REFERENCES

Akgündüz, D., Aydeniz, M., Çakmakçı, G., Çavaş, B., Çorlu, M. S., Öner, T., & Özdemir, S. (2015). *STEM eğitimi Türkiye raporu*. Istanbul: Scala Publication.

Atkinson, R., & Mayo, M. (2010, December). *Refueling the U.S. innovation economy fresh approaches to science, technology, engineering and mathematics (STEM) education*. The Information Technology & Innovation Foundation, Forthcoming. Retrieved from https://ssrn.com/abstract=1722822

Bandura, A. (1999). Social cognitive theory: An agentic perspective. *Asian Journal of Social Psychology, 2*(1), 21-41.

Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York, NY: W. H. Freeman and Company.

Baran, E., Canbazoglu-Bilici, S., Mesutoglu, C. & Ocak, C. (2016). Moving STEM beyond schools: Students' perceptions about an out-of-school STEM education program. *International Journal of Education in Mathematics, Science and Technology*, *4*(1), 9-19. doi: 10.18404/ijemst.71338

Bentler, P. M., & Chou, C. P. (1987). Practical issues in structural equation modeling. *Sociological Methods and Research*, *16*(1), 78-117. doi: 10.1177/0049124187016001004

Bouffard-Bouchard, T., Parent, S., & Larivee, S. (1991). Influence of self-efficacy on self-regulation and performance among junior and senior high-school-age students. *International Journal of Behavioral Development*, *14*(2), 153-164. doi: 10.1177/016502549101400203

Breiner, J. M., Harkness, S. S., Johnson, C. C., & Koehler, C. M. (2012). What is STEM? A discussion about conceptions of STEM in education and partnerships. *School Science and Mathematics, 112*(1), 3-11. doi: 10.1111/j.1949-8594.2011.00109.x

Cai, L., Thissen, D., & du Toit, S. H. C. (2017). IRTPRO 4.2 for Windows [Computer software]. Skokie, IL: Scientific Software International.

Chen, F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modelling*, *14*(3), 464-504. doi: 10.1080/10705510701301834

Chen, Y. F., Cannady, M. A., Schunn, C., & Dorph, R. (2017) *Measures technical brief: Competency beliefs in STEM*. Retrieved from: http://activationlab.org/wp-content/uploads/2018/03/CompetencyBeliefs_STEM-Report_20170403.pdf

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*(2), 233-255. doi: 10.1207/S15328007SEM0902_5

Chittum, J. R., Jones, B. D., Akalin, S., & Schram, Á. B. (2017). The effects of an afterschool STEM program on students' motivation and engagement. *International Journal of STEM education*, *4*(11), 1-16. doi: 10.1186/s40594-017-0065-4

Chung, J., Cannady, M. A., Schunn, C., Dorph, R., & Vincent-Ruz, P. (2016). *Measures technical brief: Competency beliefs in science.* Retrieved from: http://activationlab.org/wp-content/uploads/2018/03/Competency-Beliefs-Report-3.2-20160331.pdf

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New York, NY: Routledge.

Colakoglu, M. H. (2016). STEM applications in Turkish science high schools. *Journal of Education in Science, Environment and Health (JESEH), 2*(2), 176-187.

Colakoğlu, M. H., & Gökben, A. G. (2017). Türkiye'de eğitim fakültelerinde fetemm (stem) çalışmaları. *İnformal Ortamlarda Araştırmalar Dergisi*, *2*(2), 46-69.

Corlu, M. S., Capraro, R. M., & Capraro, M. M. (2014). Introducing STEM education: Implications for educating our teachers in the age of innovation. *Eğitim ve Bilim*, *39*(171), 74-85.

Dawes, M. E., Horan, J. J., & Hackett, G. (2000). Experimental evaluation of self-efficacy treatment on technical/scientific career outcomes. *British Journal of Guidance & Counselling*, *28*(1), 87-99. doi: 10.1080/030698800109637

Durik, A. M., Vida, M., & Eccles, J. S. (2006). Task values and ability beliefs as predictors of high school literacy choices: A developmental analysis. *Journal of Educational Psychology*, *98*(2), 382-393. doi: 10.1037/0022-0663.98.2.382

English, L. D. (2016). STEM education K-12: Perspectives on integration. *International Journal of STEM Education*, *3*(3), 1-8. doi: 10.1186/s40594-016-0036-1

Ercan, S., Altan, E. B., Taştan, B., & Dağ, İ. (2016). Integrating GIS into science classes to handle STEM education. *Journal of Turkish Science Education*, *13*, 30-43. doi: 10.12973/tused.10169a

Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. London: Sage.

Gainor, K. A. (2006). Twenty-five years of self-efficacy in career assessment and practice. *Journal of Career Assessment*, *14*(1), 161-178. doi: 10.1177/1069072705282435

George, D., & Mallery, P. (2001). *SPSS for Windows Step by Step: A Simple Guide and Reference*. Boston: Allyn & Bacon.

Green, A., & Sanderson, D. (2018). The roots of STEM achievement: An analysis of persistence and attainment in STEM majors. *The American Economist*, *63*(1), 79-93. doi: 10.1177/0569434517721770

Hacıoğlu, Y., Yamak, H., & Kavak, N. (2016). Mühendislik tasarim temelli fen egitimi ile ilgili ögretmen görüsleri. *Bartin Üniversitesi Egitim Fakültesi Dergisi*, *5*(3), 807-830. doi: 10.14686/buefad.v5i3.5000195411

Hackett, G., & Betz, N. (1982, March). *Mathematics self-efficacy expectations, math performance, and the consideration of math-related majors*. Paper presented at the annual meeting of the American Educational Research Association. New York.

Hambleton, R. K., & Jones, R. W. (1993). An NCME instructional module on: Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, *12*(3), 38-47. doi: 10.1111/j.1745-3992.1993.tb00543.x

Hambleton, R. K., Merenda, P. F., & Spielberger, C. D. (2012). *Adapting educational and psychological tests for cross-cultural assessment*. New York, NY: Psychology Press.

Han, S., Capraro, R. M., & Capraro, M. M. (2016). How science, technology, engineering, and mathematics project-based learning affects high-need students in the US. *Learning and Individual Differences*, *51*, 157-166. doi: 10.1016/j.lindif.2016.08.045

Han, S., Yalvac, B., Capraro, R. M., & Capraro, M. M. (2015). In-service teachers' implementation and understanding of STEM project-based learning. *Eurasia Journal of Mathematics, Science & Technology Education, 11*(1), 63-76. doi: 10.12973/eurasia.2015.1306a

Hidi, S., & Ainley, M. (2008). Interest and self-regulation: Relationships between two variables that influence learning. In D. H. Schunk & B. J. Zimmerman (Eds.), *Motivation and self-regulated learning: Theory, research, and applications* (pp. 77-109). Lawrence Erlbaum Associates Publishers.

International Test Commission. (2017). The ITC guidelines for translating and adapting tests (2nd ed.). Retrieved from https://www.intestcom.org/files/guideline_test_adaptation_2ed.pdf

Jenson, R. J., Petri, A. N., Day, A. D., Truman, K. Z., & Duffy, K. (2011). Perceptions of self-efficacy among STEM students with disabilities. *Journal of Postsecondary Education and Disability*, *24*(4), 269-283.

Jinks, J., & Lorsbach, A. (2003). Introduction: Motivation and self-efficacy belief. *Reading and Writing Quarterly, 19*(2), 113-118. doi: 10.1080/10573560308218

John, M., Bettye, S., Ezra, T., & Robert, W. (2016). A formative evaluation of a southeast high school integrative science, technology, engineering, and mathematics (STEM) academy. *Technology in Society*, *45*, 34-39. doi: 10.1016/j.techsoc.2016.02.001

Johnson, C. C., Peters-Burton, E. E., & Moore, T. J. (2016). *STEM road map: A framework for integrated STEM education*. New York, NY: Routledge, Taylor & Francis Group.

Kanny, M. A., Sax, L. J., & Riggers-Piehl, T. A. (2014). Investigating forty years of STEM research: How explanations for the gender gap have evolved over time. *Journal of Women and Minorities in Science and Engineering*, *20*(2), 127-148. doi: 10.1615/JWomenMinorScienEng.2014007246

Lent, R. W., Brown, S. D., & Larkin, K. C. (1986). Self-efficacy in the prediction of academic performance and perceived career options. *Journal of Counseling Psychology*, *33*(3), 265-269. doi: 10.1037/0022-0167.33.3.265

Milfont, T. L., & Fischer, R. (2010). Testing measurement invariance across groups: Applications in crosscultural research. *International Journal of Psychological Research, 3*(1), 111-121.

Milner, D. I., Horan, J. J., & Tracey, T. J. (2014). Development and evaluation of STEM interest and self-efficacy tests. *Journal of Career Assessment, 22*(4), 642-653. doi: 10.1177/1069072713515427

Ministry of National Education. (2009). *Ortaöğretim okulları öğrenci kulüp faliyetlerine yönelik eğitim materyali ve donanım ihtiyacının değerlendirilmesi*. Ankara: EARGED.

Ministry of National Education. (2016). *STEM Education Report*. Ankara: SESAM.

Ministry of National Education. (2018a). *Fen bilimleri dersi öğretim programı*. Ankara: MEB.

Ministry of National Education. (2018b). *Matematik dersi öğretim programı*. Ankara: MEB.

Monterastelli, T., Bayles, T., & Ross, J. (2008, June). *High school outreach program: Attracting young ladies with "engineering in health care."* Paper presented at the Annual Conference, & Exposition, Pittsburgh, Pennsylvania. Retrieved from https://peer.asee.org/3621

Muthén, L. K., & Muthén, B. O. (2015). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén and Muthén.

National Research Council. (2014). *STEM integration in K-12 education: Status, prospects, and an agenda for research*. Washington, DC: National Academic Press.

Nelson, B. C., & Ketelhut, D. J. (2008). Exploring embedded guidance and self-efficacy in educational multi-user virtual environments. *Computer-Supported Collaborative Learning, 3*(4), 413-427. doi: 10.1007/s11412-008-9049-1

Pajares, F., & Miller, M. D. (1997). Mathematics self-efficacy and mathematical problem solving: Implications of using different forms of assessment. *The Journal of Experimental Education, 65*(3), 213-228. doi: 10.1080/00220973.1997.9943455

Reckase, M. D. (2009). *Multidimensional item response theory.* New York, NY: Springer. doi: 10.1007/978-0-387-89976-3

Sireci, S. G., & Sukin, T. (2013). Test validity. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I. C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds.), *APA handbooks in psychology. APA handbook of testing and assessment in psychology, Vol. 1. Test theory and testing and assessment in industrial and organizational psychology* (pp. 61-84). Washington, DC: American Psychological Association. doi: 10.1037/14047-004

Smith, S. M. (2019). *A comparison of computer-based and robotic programming instruction: Impact of Scratch versus Cozmo on middle school students' computational thinking, spatial skills, competency beliefs, and engagement* (Doctoral dissertation, Kent State University).

Tellhed, U., Backström, M., & Björklund, F. (2017). Will I fit in and do well? The importance of social belongingness and self-efficacy for explaining gender differences in interest in STEM and HEED majors. *Sex Roles, 77*(1-2), 86-96. doi: 10.1007/s11199-016-0694-y

Tsupros, N., R. Kohler, and J. Hallinen (2009). *STEM education: A project to identify the missing components*. Intermediate Unit 1 and Carnegie Mellon, Pennsylvania.

TÜSİAD (2019). *2023'e doğru Türkiye'de STEM gereksinimi*. https://tusiad.org/tr/yayinlar/raporlar/item/9735-2023-e-dog-ru-tu-rkiye-de-stem-gereksinimi adresinden edinilmiştir.

Ullman, J. B. (2001). Structural equation modeling. In B. Tabachnick & L. S. Fidell (Eds.), *Using multivariate statistics* (4th ed.) (pp. 653-771). Boston, MA: Allyn & Bacon.

Urdan, T. C. (2010). *Statistics in plain English* (3rd ed.). New York, NY: Taylor & Francis Group.

Van de Vijver, F., & Hambleton, R. K. (1996). Translating tests: Some practical guidelines. *European Psychologist, 1*(2), 89-99.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices and recommendations for organizational research. *Organizational Research Methods, 3*(1), 4-70. doi: 10.1177/109442810031002

Vincent- Ruz, P., & Schunn, C. D. (2017). The increasingly important role of science competency beliefs for science learning in girls. *Journal of Research in Science Teaching*, *54*(6), 790-822. doi: 10.1002/tea.21387

Yerdelen, S., Kahraman, N., & Taş, Y. (2016). Low socioeconomic status students' STEM career interest in relation to gender, grade level, and STEM attitude [Special issue]. *Journal of Turkish Science Education (TUSED), 13*, 59-74.

Zeldin, A. L., Britner, S. L., & Pajares, F. (2008). A comparative study of the self- efficacy beliefs of successful men and women in mathematics, science, and technology careers. *Journal of Research in Science Teaching*, *45*(9), 1036-1058. doi: 10.1002/tea.20195

Zimmerman, B. J. (1995). Self-efficacy and educational development. In A. Bandura, (Ed.), *Self-efficacy in changing societies* (pp. 202-231). New York, NY: Cambridge University Press.

Zimmerman, B. J. (2000). Self-efficacy: An essential motive to learn. *Contemporary Educational Psychology*, *25*(1), 82-91. doi: 10.1006/ceps.1999.1016

Bilim, Teknoloji, Matematik ve Mühendislik Alanlarında Öz Yeterlik İnanç Ölçeğinin Türkçe'ye Uyarlaması ve Ölçme Değişmezliğinin Test Edilmesi

Giriş

Bilim, Teknoloji, Mühendislik ve Matematik alanları (STEM) eğitimi bu alanların bir bütün olarak ele alınması ile günlük yaşam problemlerinin çözümü ile ilgilenmektedir (Breiner, Harkness, Johnson, & Koehler, 2012; Johnson, Peters-Burton, & Moore, 2016; National Research Council-NRC, 2014; Tsupros, Kohler, & Hallinen, 2009). Türkiye de STEM eğitimine önem veren ve bu konuda yatırım yapan ülkelerdendir (Akgündüz ve diğerleri, 2015; TÜSİAD, 2019). Öz yeterlik inançları akademik başarıda önemli rol oynayan faktörlerden birisidir (Kanny, Sax & Riggers-Piehl, 2014). Ayrıca araştırmalar öz yeterlik inançları ve ilgi arasında pozitif bir ilişki göstermektedir (Hidi & Ainley, 2008). Bunun yanı sıra bireyler mesleki tercihlerini yaparken başarılı olacaklarını düşündükleri alanları tercih etmektedirler (Durik, Vida, & Eccles 2006; Gainor, 2006). Bu sebeple STEM eğitimi çerçevesinde öğrencilerin öz yeterlik inançlarını ölçerek STEM eğitimi ile ilişkilendirmek önemlidir. Ancak, Türkiye'de STEM öz yeterlik becerilerini ile ilgili bir ölçek bulunmamaktadır. Chen, Cannady, Schunn ve Dorph (2017) İngilizce olarak STEM yeterlik inançları ölçeği geliştirmiştir. Bu çalışma da bu ölçeğin Türkçe'ye uyarlamasını yapmayı amaçlamaktadır. Bu ölçeğin Türkçe'ye kazandırılmasının Türkiye'deki STEM çalışmalarına katkı sağlaması beklenmektedir. Bu çalışmanın iki ana amacı bulunmaktadır. Birinci amaç ölçeğin uyarlanarak yapısının Türk öğrencilerden toplanan veri ile test edilmesidir. İkinci amaç ise yapının cinsiyet grupları ve STEM ile ilgili kariyer hedefi olan ve olmayan öğrenci grupları arasında ölçme değişmezliği gösterip göstermediğinin incelenmesidir. Ayrıca öğrencilerin ölçekte elde edilen puanları cinsiyet, okul türü ve kariyer hedefleri değişkenleri bakımından karşılaştırılmıştır. Bu çalışmanın araştırma soruları aşağıdaki gibidir.

1) STEM yeterlik inançları ölçeğinin orijinal yapısı Türk öğrencilerinin verisi ile desteklenmekte midir?
2) Elde edilen yapı kızlar ve erkekler için ölçme değişmezliği göstermekte midir?
3) Elde edilen yapı STEM ile ilgili kariyer hedefi olan ve olmayan öğrenciler için ölçme değişmezliği göstermekte midir?
4) Öğrencilerin ölçekte elde edilen puanları cinsiyet, okul türü ve kariyer hedefleri değişkenleri bakımından farklılık göstermekte midir?

Yöntem

Örneklem

Uyarlama aşaması pilot uygulama ve asıl uygulama basamaklarından oluşmuştur. Pilot uygulamaya 77 ortaokul öğrencisi, asıl uygulamaya 330 ortaokul öğrencisi katılmıştır. Asıl uygulamada kız ve erkek sayıları birbirine yakındır. Öğrencilerin %92'si devlet okulu, %8'i ise özel okul öğrencisidir.

Ölçme aracı

Ortaokul öğrencilerinin STEM yeterlik inançlarını ölçmeyi amaçlayan bu ölçek 12 maddeden oluşmakta ve 4'lü Likert tipi yapıya sahiptir. Ölçme aracı "Sınıfta sorulan matematik sorularını çözebilirim" ve "Evimdeki teknoloji uzmanı benim" gibi maddelerden oluşmaktadır.

Veri analizi

Veri analizi kısmında uyarlama aşamaları, pilot çalışma, güvenirlik ve geçerlik analizleri ve ölçme değişmezliği analizleri ile ilgili yapılanlar açıklanmaktadır.

Uyarlama aşamasında ilk olarak gerekli izinler alınmıştır. Ardından bu konuda tecrübeli uzmanlar tarafından ölçeğin çevirisi gerçekleştirilmiştir. Bağımsız yapılan bu çeviri işleminden sonra araştırma ekibinin de sürece dahil olması ile bu aşama tamamlanmıştır. Ardından geri çeviri aşaması gerçekleştirilmiştir. Son aşama olarak ölçeğin Türkçesi uzmanlar tarafından incelenmiştir. Araştırma ekibi ise gerekli kontrolleri yapmıştır. Pilot aşamasında ifadelerin anlaşılırlığı incelenmiş ve gerekli düzeltmeler yapılmıştır.

Güvenirlik için Cronbach Alfa iç tutarlılık katsayısı hesaplanmıştır. Bu değerin .70'ten büyük olması beklenmektedir. Ayrıca, madde bazında sorunları görebilmek için düzeltilmiş madde toplam korelasyonu hesaplanarak değeri .30 altında olan maddeler incelemeye alınmıştır.

Geçerlik çalışmaları için doğrulayıcı faktör analizi yapılmıştır. Doğrulayıcı faktör analizinde daha önce belirlenmiş olan bir yapının toplanan verilerle uyumu incelenir. CFI (Comparative Fit Index), TLI (Tucker Lewis index) ve RMSEA (Root Mean Square Error of Approximation) gibi örneklem sayısından direk etkilenmeyen uyum değerleri incelenerek testin yapısı test edilmektedir. CFI ve TLI değerlerinin .95'ten büyük, RMSEA değerinin ise .06'dan küçük olması istenmektedir (Ullman, 2001). Orijinal ölçekte belirlenen tek faktörlü yapı doğrulayıcı faktör analizi kapsamında test edilmiştir.

Gruplar arası karşılaştırma akademik çalışmalarda önemli bir yer tutmaktadır. Ancak bu karşılaştırmaların yapılabilmesi için ölçülen kavramların alt gruplar için aynı anlam taşıyıp taşımadığı test edilmelidir. Bu sebeple yapısal, metrik ve skalar değişmezlik incelenmiştir. Yapısal modelde gruplar için yapı benzerliğine, metrik modelde faktör yüklerinin eşitliğine, skalar modelde ise ortalamaların eşitliğine bakılmıştır. Modeller arası uyum değeri farkının CFI için .01'den RMSEA için .015'ten küçük olması istenir (Chen, 2007; Cheung & Rensvold, 2002). Öğrenci puanları ise çok boyutlu madde tepki kuramı kullanılarak kestirilmiştir.

Sonuç ve Tartışma

İç tutarlılık

12 maddeden oluşan ölçeğin Cronbach Alfa iç tutarlılık değeri .83 olarak hesaplanmıştır. Bu değer ölçeğin iyi düzeyde iç tutarlılığa sahip olduğunu göstermektedir. Düzeltilmiş madde-toplam korelasyon değerlerinin hepsinin .30 değerinin üzerinde olması ise madde bazında bir problem olmadığını göstermektedir.

Doğrulayıcı faktör analizi

Doğrulayıcı Faktör Analizi (DFA) sonuçlarına göre elde edilen veri tek faktörlü yapıyı desteklememektedir (CFI = .890 < .95; TLI = .866 < .95; RMSEA = .117 > .06). Bu sebeple Açımlayıcı Faktör Analizi (AFA) yapılarak faktör yapısı incelenmiştir. AFA sonuçları ölçekteki maddelerin 2 farklı boyut oluşturduklarını göstermektedir. Madde 1, 8 ve 9 ayrı bir faktör ile ilişkidirler. Bu maddeler incelendiğinde bu maddelerin matematik ile ilgili oldukları diğer maddelerin ise bilim, teknoloji ve mühendislik ile ilgili oldukları görülmektedir. Bu faktörlere Mat ve STE isimleri

verilmiştir. Burada ortaya çıkan iki boyutlu yapı DFA ile incelendiğinde ise yapının veri tarafından doğrulandığı görülmektedir (CFI = .974 > .95; TLI = .968 > .95; RMSEA = .057 < .06). Bu sebeple ölçeğin Türkçe uyarlamasının iki boyutlu bir yapıya sahip olduğuna karar verilmiştir.

Ölçme değişmezliği

Kızlar ve erkeklerden elde edilen veri yapı değişmezliğini desteklemektedir (TLI = .971, CFI = .975, RMSEA = .058). Faktör yüklerinin eşitlenmesi ile elde edilen model karşılaştırması da metrik değişmezliğin olduğunu ortaya koymaktadır (ΔCFI = .001, ΔRMSEA = -.003). Ancak, skalar değişmezlik verilerinde ΔRMSEA değeri iyi iken ΔCFI değeri istenen seviyede değildir (ΔCFI = -.016, ΔRMSEA = .006). Modifikasyon değerleri bu sorunun 7. maddeden kaynaklanabileceğini göstermektedir. Bu madde üzerindeki sınırlılıklar kaldırıldığında ise elde edilen değerler skalar değişmezliğin de desteklendiğini göstermektedir (ΔCFI = -.010, ΔRMSEA = .002). Bu sebeple madde 7 dışında testin ölçme değişmezliğine sahip olduğu ve kızlar ve erkeklerin puanlarını karşılaştırmada kullanılabileceği sonucuna ulaşılmıştır. Madde 7 "Fikrimi söylerken kanıtlar sunmakta iyiyim" ifadesinden oluşmaktadır.

Kariyer hedefleri STEM ile ilgili olan ve olmayan öğrenciler için de ölçme değişmezliği test edilmiştir. Kariyer hedefi grupları için elde edilen veri yapı değişmezliğini desteklemektedir (TLI = .961, CFI = .969, RMSEA = .063). Faktör yüklerinin eşitlenmesi ile elde edilen model karşılaştırması da metrik değişmezliğin olduğunu ortaya koymaktadır (ΔCFI = .002, ΔRMSEA = .005). Skalar değişmezlik verilerinde de ΔCFI değeri ve ΔRMSEA değeri beklenen düzeydedir (ΔCFI = .000, ΔRMSEA = .009). Bu bulgular kariyer grup ortalamalarının karşılaştırılabileceğini göstermektedir.

Grupların karşılaştırılması

Öğrencilerin çok boyutlu madde tepki kuramı kullanılarak kestirilen mat ve STE puanları cinsiyet, okul türü ve kariyer hedefleri değişkenleri bakımından karşılaştırılmıştır. Kızlar ve erkekler arasında istatistiksel olarak anlamlı bir fark bulunamamıştır. Özel okullardaki öğrencilerin devlet okullarındaki öğrencilere göre öz yeterlik inanç puanlarının daha yüksek olduğu görülmüştür. İleride STEM ile ilgili alanlarda bir meslek sahibi olmak isteyen öğrencilerin puanları STEM dışında mesleklere yönelmek isteyen öğrencilerin puanlarından daha yüksektir. Bu sonuçlar etki büyüklüğü hesapları tarafından da doğrulanmaktadır.

Bu çalışma STEM öz yeterlik inançları ölçeğini Türkçe'ye uyarlaması bakımından önemli bir çalışmadır. Ölçeğin güvenirliği ve geçerliği ile ilgili kanıtlar sunulmuş, STEM araştırmalarında kullanılabilecek bir uyarlama olduğu ortaya konmuştur. Karşılaştırma çalışmalarında bir önkoşul olan ölçme değişmezliğinin test edilmesi ve örneklendirilmesi de önemlidir.

Elde edilen veriler uyarlanan ölçeğin faktör yapısının orijinal ölçeğin faktör yapısından farklı olduğunu göstermiştir. Bu durumun Türkiye'de STEM kavramlarının bir bütün olarak görülmemesinden kaynaklandığı düşünülmektedir. Öğretim programlarındaki vurgunun da bir bütün oluşturmadığı görülmektedir.

Özel okullardaki öğrencilerin ve STEM ile ilgili bir kariyer isteyen öğrencilerin daha yüksek STEM öz yeterlik inanç puanına sahip olmaları geçerlik için ayrıca bir kanıt olduğu düşünülmektedir. Bu ölçeğin puanları farklı öğrenci grupları için farklılık gösterebilmektedir. Özel okullarda sağlanan STEM imkanları ve uygulamaları ile devlet okullarının kısıtlı imkanları öğrencilerin öz yeterliklerinin ayrışmasına sebep olmuş olabilir. STEM ile ilgili kariyer hedefleyen öğrenciler ile farklı alanlara yönelmek isteyen öğrencilerin öz yeterlik puanlarının farklı çıkması da bu ölçeğin geçerliğini desteklemektedir.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2020; 11(2); 180-198

Revisiting Quick Big Five Personality Test: Testing Measurement Invariance across Gender

Devrim ERDEM *

Abstract
Personality is a subject that has been studied because of the social, economic, individual, and educational implications of personality. The widely used model for measuring personality is the Five-Factor Model (FFM). The robustness of the factor structure of the FFM of personality has been provided among cultures and diverse samples. The measurement tools are used to identify differences between individuals or groups. However, in order to make meaningful comparisons, it is necessary to provide the measurement equivalence among the comparison groups. Thus the current study aimed to test the measurement invariance of the Quick Big Five (QBF) items that are used in many disciplines in Turkey. For this purpose, the QBF items were investigated in terms of configural, metric, scalar and strict invariance across gender. In this research, 1114 university students aged between 17-32 years were included in the sample. Firstly, several CFAs were performed for the whole sample and then both men and women separately. The findings of the CFA revealed that the QBF model fit the data. In addition, each of the 30 items of the scale was embedded into a related latent factor in both gender groups. Secondly, sequential multiple group CFA tests to examine measurement invariance were conducted. According to the findings, full configural, partial metric and scalar invariance were fulfilled across gender. However, strict invariance could not be achieved. Imaginative and inquisitive under the openness factor were determined to cause measurement non-invariance. In conclusion, latent mean comparisons can be made by excluding these two items across gender.

Key Words: Five-factor model, personality traits, partial metric invariance, early adulthood, sex.

INTRODUCTION

Personality traits are comparatively long-lasting molds of opinions, emotions, and manners that make individuals different from each other (Bleidorn, Hopwood, & Lucas, 2018). The development of personality traits throughout the life span has been an intriguing subject. Caspi and Shiner (2006) noted that one of the important reasons for this is that there are many theoretical and practical implications and outcomes of understanding personality development (cited in Morizot, 2014). Perhaps the most popular personality conceptualization used in personality measurement is the Five-Factor Model (FFM). This model arranges personality into five trait domains. However, this classification does not mean that all personality traits can be reduced into five factors; rather, the "big five" should be seen as broad but comprehensive factors based on a series of associated items (Mueller & Plug, 2006; Paunonen & Ashton, 2001). Almost universally, researchers have reached a consensus on the representation of the Five-Factor Personality Model (John, Neumann, & Soto, 2008; Korkmaz, Somer, & Gungor, 2013; McCrae, Terracciano, & Pro, 2005).

The theoretical foundations of the Five-Factor Model (FFM) were formed by the lexical hypothesis (Allport & Odbert, 1936 as cited in Poropat, 2009). According to this hypothesis, the most prominent features of people as personality traits eventually become part of their own language and show themselves in the language they use. Based on this hypothesis, it was envisioned that personality traits could be identified by looking at the descriptive adjectives in languages. Adjectives that may be indicative of personality, especially in English, have been determined. Afterwards, it was possible to develop scales based on Five-Factor Model and examines their validity with factor analytical studies in other languages (Saucier & Goldberg, 1996).

* Assist. Prof., Niğde Ömer Halisdemir University, Faculty of Education, Niğde-Turkey, erdem_devrim@yahoo.com, ORCID ID: 0000-0003-1810-2454

To cite this article:
Erdem, D. (2020). Revisiting quick big five personality test: Testing measurement invariance across gender. *Journal of Measurement and Evaluation in Education and Psychology, 11*(2), 180-198. doi: 10.21031/epod.675796
Received: 16.01.2020
Accepted: 09.05.2020

The Big Five dimensions consist of agreeableness, extraversion, conscientiousness, neuroticism, and openness to experience. Individuals with a higher orientation in the Agreeableness dimension are known as compassionate, polite, tolerant, open to co-operation, and willing to help. Conscientiousness represents individual differences in target orientation, organized, self-discipline, impulse control, and compliance with social norms and rules. Individuals with a higher orientation in the Neuroticism dimension are considered worried, self-conscious, acting without forethought, and downbeat. They feel vulnerable, tend to experience low self-worth, and experience negative emotions relatively easily. Extraversion reflects being socially confident, willing to make friends, assertive and energetic. Individuals with a higher orientation in the Openness dimension are known as willing to try new things, broad-minded, intellectual curiosity, high imagination, creative, and artistic sensitivity (Barrick & Mount, 1991).

It is noteworthy that most of the research on personality development focuses on early adulthood (Durbin et al., 2016; Fadjukoff, Feldt, Kokko, & Pulkkinen, 2019; Johnson, Cohen, Brown, Smailes, & Bernstein, 1999; Shiner, Allen, & Masten, 2017; Soto, 2016). Longitudinal studies on the Big Five have shown that relatively great and resistant changes in personality have occurred in early adulthood (Roberts, Walton, & Viechtbauer, 2006). This could be due to the fact that "important biological, social, and psychological changes occur throughout childhood and adolescence" (Soto, 2016, p. 410). Hence the period from late childhood through early adulthood is called a critical personality development period (Durbin et al., 2016). Besides, the frontal lobe of the brain continues to develop until the age of 25 or 28. Further maturation of these regions of the brain enhances persons' capacity for better judgment, self-regulation, planned behaviors, and for more complex cognitive functioning. These functions do, in turn, contribute to the various developmental tasks of this age group. In addition, the period between the ages of 18 and 30 constitutes the transition to adulthood is an important stage of development in terms of sincerity, entrepreneurship, social interests, identity, work and parenting (Arnett, 2000). Indeed, research has shown that in early adulthood, interests are crystallized and balanced, and professional aspirations and prospects are delineated with more precision (Low & Rounds, 2007). Therefore researchers still have an ongoing interest in this developmental period. Moreover, personality traits are part of the individual's productivity, and it is important to examine these traits as they are directly social and economic value.

Gender Differences in Personality

Personality traits are broad and relatively stable individual differences that affect human behavior and choices. Gender differences in personality traits have always been of interest to researchers (Kajonius & Johnson, 2018). There are several reasons for this interest. First, gender differences in personality were observed in all cross-cultural studies (e.g., Costa, Terracciano, & McCrae, 2001; Guimond, 2008; Schmitt, Realo, Voracek, & Allik, 2008). It is a universal issue. Also, there is ample evidence that gender differences in personality are relatively stable throughout life (Donnellan, Conger, & Burzette, 2007). In addition, many social choices such as occupational, educational, spousal selection, conflict, and relationship regulation are related to personality (Berings, De Fruyt, & Bouwen, 2004; Bono, Boles, Judge, & Lauver, 2002; Figueredo, Sefcek, & Jones, 2006; Gasser, Larson, & Borgen, 2007). For example, although there is an increase in women's level of education and participation in "high-status professional fields, women and men are still concentrated in different occupations and educational programs, and women are still under-represented in the fields associated with physical science, engineering, and applied mathematics" (Eccles, 2011, p. 195). Unfortunately, there still exists a large gender aperture in mathematics, technology, engineering, and science majors (Cole & Espinoza, 2008; Langen & Dekkers, 2005; Legewie & DiPrete, 2014; Wang & Degol, 2017). Thus, it may be possible to monitor and improve the development of individuals, especially of women, in terms of education, skills and occupations by examining psychological factors such as personality traits, of course, along with various social policies toward gender equality.

Meta-analytic studies have shown that gender differences in psychological variables vary according to the construct examined. For example, men dominate sexual and physical aggression, status-seeking, and risk-taking behavior (Buss, 2004; Lynn, 1993). In contrast, devotion, care and benevolence

tendencies are higher among women in all societies (Browne, 2006). The effect of personality on earnings (income) of women and men is also noteworthy. Compatibleness appears to be higher in women and lower in men and functions as a factor for women to consent to lower wages (Mueller & Plug, 2006). Similarly, agreeableness and neuroticism consistently emerge as two traits that show the highest gender differences in women (Bouchard & Loehlin, 2001; Costa et al., 2001; Kajonius & Johnson, 2018). Self-identity and self-esteem are associated with sensitivity to others and focusing on relationships in women; in contrast, in men, it is associated with a tendency to establish autonomy and ascendancy over others (Josephs, Markus, & Tafarodi, 1992). The FFM suggests that gender differences are usually small or moderate but significant, in terms of the effect size, and that men tend to show greater differences in personality traits than women (Borkenau, McCrae, & Terracciano, 2013; Lippa, 2010).

On the other hand, the literature review shows that the last two decades has added a new perspective to the results of research on personality and gender. Surprisingly, more gender-based differences have been reported in more gender-egalitarian societies (Fischer & Manstead, 2000; Kajonius & Johnson, 2018; Schmitt, Long, McPhearson, O'Brien, Remmert, & Shah, 2017). In other words, gender differences in personality are greater in more individual, more economically developed and more egalitarian societies, because this like of conditions lets men and women to more freely express their intrinsic dispositions (Falk & Hermle, 2018). Therefore, such studies are crucial in order to grasp the origin of gender distinctions in personality traits and to broaden our understanding of this issue.

Personality and Academic Performance

Personality and its relations with social and economic structures have always been a lively research topic (Funder, 2001). On the other hand, the impact of personality on academic achievement and its educational implications have been ignored until the last decades. As Poropat (2014) pointed out that "One of the areas in which both educators and learners have been under-informed is the role of individual differences in learning and education, especially with respect to temperament and personality" (p. 24). Personality keeps a substantial role in students' school experience and academic success (Matthews, Zeidner, & Roberts, 2006). The desire for performance in a job or academic activity and continuity in performance was found more decisive than FFM factors rather than mental ability (Heckman, Stixrud, & Urzua, 2006, Judge & Ilies, 2002; Willingham, Pollack, & Lewis, 2002). Non-mental skills function a major role in the school performance of children and adolescents (Duckworth & Seligman, 2005; Matthews et al., 2006). Some studies have shown that personality traits predict academic achievement better than indicators of cognitive measures (Lounsbury, Sundstrom, Loveland, & Gibson, 2003).

Motivation, which has an important function in learning, is conceptualized as a personality trait (Rindermann & Neubauer, 2001). Conscientiousness has been identified as the strongest dimension of FFM in predicting academic performance (Chamorro-Premuzic, & Furnham, 2003; Dumfart & Neubauer, 2016; Nguyen, Allen, & Fraccastoro, 2005; O'Connor & Paunonen, 2007; Poropat, 2009). Similarly Noftle and Robins (2007) pointed out Conscientiousness was the most powerful predictor of both high school and college GPA. Emotional stability (low neuroticism) is related to self-efficacy (Judge, Erez, Bono, & Thoresen, 2002) and predicts academic achievement (Poropat, 2011). Noftle and Robins (2007) found Openness was the most potent predictor of SAT verbal scores. Openness to experience has been associated with learning, motivation for learning, intelligence, critical thinking, and lexical intellect (Bidjerano & Dai, 2007; De Raad & Schouwenburg, 1996; Klein & Lee, 2006). Obviously, it is substantial to investigate the academic performance of individuals because significant investments are made in education by communities and individuals indicating the high worth given to educational performance (Poropat, 2009). The strong relationships between academic performance and Big Five personality factors indicate that we need to focus more on personality traits in terms of education.

Measurement Invariance on Big Five

Empirical studies with different cultures and settings supported the robustness and generalizability of the Big Five personality factor structure (John & Srivastava, 1999). In addition, there is considerable evidence that the Big Five personality traits have predictive validity in childhood, adolescence, and adulthood, as well as repeatability of factor structure during different developmental periods (see in Morizot, 2014). However, in order to interpret the differences or similarities between the comparison groups of a psychological construct, it is necessary to test the invariance of the psychological construct through measurement invariance. As mentioned so far, investigating personality traits is crucial to provide an understanding of educational decisions and developmental screening. Although there are significant differences between males and females, studies showing the equivalence of factor structures at the latent mean level are too limited in personality research (Morizot, 2014; Samuel, South, & Griffin, 2015). Therefore, there is a need for research that supports the structure of the Big Five, which is widely used in almost every discipline (psychology, health, economy, education, sociology, etc.) with further validity analyzes. If the scalar measurement invariance can be achieved in comparison groups for Big Five construct, it is possible to make meaningful comparisons between the latent means (Ock, McAbee, Mulfinger, & Oswald, 2020; Sass, 2011). Otherwise, it cannot be determined whether the resulting differences can actually be attributed to the true difference between the groups or to a situation stemming from the lack of equivalence of the psychological construct. In this case, both the validity and generalizability of the psychological structure become problematic.

While several Turkish instruments have been developed based on the Big Five theory, they are often too long for practical applications. Also, the measurement invariance of such scales has not been studied. Only Korkmaz et al. (2013) examined the measurement invariance of gender in high school adolescents on a 200-item scale developed by them. However, further research is needed with various developmental groups. In behavioral sciences, researchers tend to view scales that are above 40 items as "substantial length" and generally prefer "abridged" versions (Roets & Van Hiel, 2011). Since the Quick Big Five (QBF) scale is a relatively "brief" scale, it provides ease of use and application. Indeed, this is why it has been preferred in many research and used widely by professionals from various disciplines (education, health, economics, psychology, etc.). Understanding the development of personality traits throughout life span has theoretical and practical consequences (Roberts, Kuncel, Shiner, Caspi, & Goldberg, 2007). In particular, it is important to examine the validity of the scores obtained from relatively shorter self-report tools through further studies. Such studies are also important in contributing to current discussions about the nature of the personality and in terms of understanding cultural differences in personality factors. Only with such an evidence, the use of the current instrument in university-counseling centers for clinical use or for the use of researchers intending to make gender comparisons could yield to sound results. Therefore, the current study has two aims: (1) to test model fit of the Quick Big Five (QBF) on a Turkish early adulthood sample, and (2) to test the measurement invariance of the scale items. Concerning the second purpose, QBF-30 items under five factors were examined in terms of configural, metric, scalar and strict invariance across gender.

METHOD

This study aimed at investigating measurement invariance of the Quick Big Five scale across gender. In this section the participants, data collection tool, and the data analysis were described.

Participants

The sample was comprised of 1114 university students, aged 17–32 years (M_{age} = 20.8, $Median_{age}$ = 21, *SD* = 2.4), from Central-Anatolia Turkey. Among them were 659 females (59%) and 455 males (41%). Information on students' faculty and grade were presented in the Table 1. Data were collected during the 2018-2019 academic year.

Table 1. Participants' Faculty and Grade

		f	%
Faculty	Missing	26	2.3
	Education	270	24.2
	Science and Literature	121	10.9
	Economics and Administrative Sciences	236	21.2
	Engineering	201	18.0
	Architecture	62	5.6
	Communication	47	4.2
	Agricultural Sciences and Technologies	71	6.4
	Islamic Sciences	40	3.6
	Medicine	40	3.6
	Total	1114	100.0
Grade	Preparatory	44	3.9
	1st	219	19.7
	2nd	242	21.7
	3rd	419	37.6
	4th	190	17.1
	Total	1114	100.0

Data Collection Instrument

The QBF is a scale measuring personality traits. The QBF was adapted from Goldberg's Big Five Personality scale consisting of 100 adjectives by reducing the number of items to 30 (Vermulst & Gerris, 2005). There were two groups in their study. There were 12107 participants (5865 male) in the 12-18 age group and 7172 participants (3622 male) in the 19 and older age group. The QBF personality dimensions are extraversion, conscientiousness, agreeableness, emotional stability, and openness. Each personality trait is measured with six items; thus, the scale consists of 30 items. The items are marked on a 7-grade rating scale that ranges from *completely untrue* (1 point) to *completely true* (7 points). The 12 items in the scale are reverse coded. The scores for each subscale range from 6 to 42. High scores indicate high levels of the relevant personality dimension. Confirmatory factor analysis (CFA) was used to determine the factor structure of the scale. CFA results showed that the 5-factor structure was confirmed (RMSEA = .05, CFI = .96). The Cronbach Alpha values for the sub-scales were .81 for extraversion, .80 for agreeableness, .86 for conscientiousness, .78 for emotional stability and .73 for openness to experience respectively. The test-retest reliability of the scale was also acceptable (Vermulst & Gerris, 2005). The validity studies of the QBF have been conducted in different adolescent and adult groups until now (e.g., Borghuis et al., 2017; Klimstra et al., 2013; Manders, Scholte, Janssens, & De Bruyn, 2006).

The QBF was adapted to Turkish culture by Morsunbul (2014). In his study, 793 participants were included consisting of two age groups: adolescent group aged 14-17 and university students aged 18-22. Based on the CFA results (χ^2/df = 3.76, GFI = .91, CFI = .92, NFI = .91, NNFI = .91 RMSEA = .08), the five-factor structure of the scale was confirmed with the Turkish sample. The Cronbach's Alpha coefficients of the subscales ranged from .71 to .81 in the adaptation study.

Before completing the QBF, participants were asked for gender, age, grade and faculty information. Informed consent was obtained from all participants involved in the study. Data for this study was collected during the academic year of 2018-2019. Although at the time of data collection the institutional ethical permission was not obtained, all necessary steps were taken to ensure the ethical rights of the participants. The nature of questions/items on the surveys was not of any sort to pose any likely distress for participating students. Nor the results of the study pose any risk for bridging of confidentiality. Thus, during data collection, in reporting the findings as well as by not obtaining or revealing students' names or other personal information, the study adhered to ethical principles at the utmost level.

Data Analysis

The suitability of the data for the analyses was examined before proceeding to the analyses. Data entry, missing value, outlier, and normality were evaluated with SPSS 22.0. LISREL9.2 was used for the confirmatory factor analysis (CFA) and multiple-group CFA for testing invariance across gender.

Confirmatory factor analysis

Confirmatory factor analysis (CFA) was performed to examine the model fit. The maximum likelihood estimation method with the covariance matrix was employed in the CFA. Because the chi-square (χ^2) statistic is sensitive to sample size, it may cause inflated chi-square values (Kline, 2011). Therefore, various fit indexes were also evaluated along with the chi-square statistic. The following criteria and indices recommended by Hu and Bentler (1999) and Kline (2011) were taken into consideration. The comparative fit index (CFI), which is less sensitive to large samples and the non-normed fit index (NNFI), which is generally considered to be relatively independent of sample size were preferred as incremental fit indexes. The goodness of fit index (GFI) and the root-mean-square error of approximation (RMSEA) were chosen as absolute fit indexes considered while assessing model fit in CFA. While "an absolute-fit index directly assesses how well an a priori model reproduces the sample data" (Hu & Bentler, 1998, p.426) "incremental fit indexes evaluate model fit by comparing a target model with a more restricted, nested baseline model" (Beauducel & Wittmann, 2005, p.45). The ratio of chi-square to degrees of freedom (χ^2/df) values less than 5 suggest sufficient fit; the CFI, GFI, and NNFI values .90 or greater indicate adequate model fit. The RMSEA values .08 or less point out a good fit.

Measurement invariance

Measurement invariance has been viewed as a way of assessing the applicability of test instruments when the same psychological construct is intended to be measured in a different group (Cheung & Rensvold, 2002). In this study, measurement invariance was tested by multiple groups confirmatory factor analysis (MGCFA). A series of successive tests are followed for the measurement invariance. First, the configural model is tested. When testing the configural invariance, factor loadings and intercepts are not restricted, except for reference indicators, and factor means are fixed at 0 for both groups (Putnick & Bornstein, 2016). Ensuring the configural invariance is a prerequisite for the metric, scalar, and strict invariance. After establishing the configural invariance, metric invariance is tested. When testing the metric invariance, the factor loadings are equalized, but intercepts are not restricted between the groups (Putnick & Bornstein, 2016). After achieving the metric invariance, scalar invariance is tested. When testing scalar invariance, factor loadings and intercepts are restricted, but error variances were allowed to vary across groups. If scalar invariance is obtained, then strict invariance is tested. When testing strict invariance (invariant uniqueness), all error variances are constrained to be equal across groups (Milfont & Fischer, 2010).

Chi-square difference test ($\Delta\chi^2$) is employed to compare these nested models (Brown, 2006; Dimitrov, 2010; Tabachnick & Fidell, 2001). The presence of a non-significant difference for each model indicates that the measurement invariance is accepted. However, if it is considered that the chi-square test is affected by the sample size, it is recommended to use another indicator. Therefore, following to recommendation of Cheung and Rensvold (2002) CFIs difference values (ΔCFI) were used to compare these nested models. In order to accept measurement invariance, the delta CFI value in each model tested must be greater than -0.01 (Cheung & Rensvold, 2002). When measurement invariance cannot be achieved, partial measurement invariance is examined. As Milfont and Fischer (2010) stated "partial measurement invariance may allow appropriate cross-group comparisons even if full measurement invariance is not obtained." (p.117).

According to Van De Schoot, Lugtig, and Hox (2012), the purpose of analyzing partial measurement invariance is to determine which loadings or intercepts differ between groups. The authors suggested following the steps to establish partial measurement invariance:

> Study the size of the loadings and/or intercepts, and constrain all loadings and intercepts, except for the one loading/intercept with the largest unstandardized difference, which is released. Subsequently, compare this new model with the old Model 1 or 2. If $\Delta\chi^2$ is now insignificant, partial invariance is established. If $\Delta\chi^2$is still significant release another item, and continue until the item that causes MI not to hold is identified. (p.491)

In line with the recommendations of these researchers, the suitability of individual parameter equality constraints was examined when it is necessary to investigate the partial invariance. In this current study, while checking partial invariance ΔCFI value along with $\Delta\chi^2$ was taken into consideration.

RESULTS

Confirmatory Factor Analysis

A CFA was conducted to investigate the model fit to the Quick Big Five scale. The fit indexes for the five-factor structure with 30 items were found for the full sample as follows (in Table 2): $\chi^2_{(395)}$ = 4457.75 ($p < .000$) and χ^2/df = 11.28 did not support the fit of the model. As already mentioned, this was an expected finding related to the sensitivity to the sample size of the chi-square statistics. The other fit indexes were found as follows: CFI = .94, NNFI = .94, GFI = .93 and RMSEA = .082 [90% lower-upper confidence interval .080 - .085]. The RMSEA deviated slightly from model fit. On the other hand, based on the values concerning CFI, NNFI, and GFI, the model-data fit was met. According to the *t*-test, factor loadings in CFA were found significant at .05 level. In light of these findings, it was concluded that the model data fit for the five-factor solution of the scale was acceptable.

Measurement Invariance Across Gender

In order to examine the measurement invariance according to gender, firstly CFA was performed separately in female and male groups. According to the χ^2/df, model fit was not attained for both the female and the male groups. However considering the alternative fit indices it was concluded that the model fit was acceptable for the female as well as the male groups based on the CFI, NNFI, and GFI values. On the other hand, RMSEA values both females and males indicated a bit model misfit. These findings presented in Table 2.

Table 2. Goodness-of-fit Indexes for the Full Sample and the Baseline Model across Gender

Group	χ^2	*df*	χ^2/df	CFI	NNFI	GFI	RMSEA	90% CI for RMSEA	
								Lower	Upper
Full	4457.75***	395	11.28	.94	.94	.93	.08	.080	.085
Female	2526.20***	395	6.4	.93	.93	.90	.09	.088	.092
Male	1978.81***	395	5.0	.95	.95	.93	.09	.089	.095

*** p < .001

After the baseline model was achieved the next step was to establish configural invariance. Although conducting individual CFAs in each group (baseline models) can test configural invariance, it is still necessary to run this step in MGCFA (Milfont & Fischer, 2010). Configural model presented at Table 4 showed adequate fit to the data, except for the chi-square statistics (χ^2/df = 6.61, CFI = .93, NNFI = .92, RMSEA = .08). These findings indicated that the factorial structure of the construct was equal across gender. Standardized factor loadings, error terms and *t*-values in the baseline (configural) model were presented in Table 3.

Next, metric invariance was examined. Findings of the fit indexes of measurement invariance were presented in Table 4. While comparing nested models, the chi-square difference test and ΔCFI values were examined. The chi-square difference between metric model and configural model was

statistically significant ($\Delta\chi^2_{(30)}$ = 1820.59, $p < .0001$) and ΔCFI = -.03 < -.01; thus indicating metric invariance was not achieved. These findings showed that factor loadings could not be accepted as equal across gender groups.

Table 3. Standardized Factor Loadings, Error Terms and t-values in the Configural Model

Items		Standardized factor loadings		Standard Error		*t*-values	
		Female	Male	Female	Male	Female	Male
Agreeableness							
5	Pleasant	.55	.69	.059	.072	15.73	16.21
10	Helpful	.60	.68	.060	.072	17.03	16.21
15	Kind	.72	.73	.056	.066	21.36	18.66
20	Cooperative	.51	.57	.072	.086	14.01	13.24
22	Agreeable	.64	.66	.064	.075	18.21	15.99
28	Sympathetic	.67	.66	.063	.073	19.35	16.35
Extraversion							
4	Reserved^R	.44	.62	.071	.087	12.02	13.54
9	Quiet^R	.60	.62	.073	.086	16.61	14.46
13	Introverted^R	.65	.73	.070	.085	18.84	17.55
18	Talkative	.23	.37	.077	.092	-5.90	-7.74
21	Bashful^R	.73	.75	.072	.085	21.42	18.65
26	Withdrawn^R	.75	.71	.073	.084	21.66	17.87
Conscientiousness							
3	Sloppy^R	.18	-.07	.084	.097	4.35	2.50
8	Careful	.57	-.64	.067	.080	15.47	-14.64
12	Organized	.76	-.86	.065	.080	23.37	-21.55
17	Prompt	.56	-.65	.074	.089	15.39	-14.83
25	Neat	.74	-.82	.071	.086	22.33	-20.41
27	Systematic	.63	-.73	.070	.086	17.96	-17.18
Neuroticism							
2	Irritable^R	.39	.46	.077	.092	10.14	9.87
7	High-strung^R	.58	.62	.068	.082	16.01	14.21
11	Touchy^R	.59	.56	.074	.087	16.03	13.02
16	Anxious^R	.74	.72	.068	.079	21.72	18.03
24	Fearful^R	.62	.49	.077	.088	16.59	11.42
29	Nervous^R	.73	.69	.070	.082	21.05	17.03
Openness							
1	Imaginative*	.58	.86	.056	.070	18.29	20.47
6	Inquisitive*	.61	.84	.063	.074	18.39	20.43
14	Sophisticated	.67	.81	.056	.070	20.31	19.99
19	Innovative	.68	.80	.064	.076	20.36	20.04
23	Artistic	.57	.56	.078	.093	15.67	13.11
30	Creative	.72	.80	.064	.075	21.73	20.57

R Revised items, * non-invariance items

Partial metric invariance was investigated in order to determine which item or item groups had different factor loadings. When full metric invariance is not attained, the non-invariant items can be found by gradually releasing the factor loadings according to items with the highest modification index until a final partial metric invariance model is achieved (Cooper, Gomez, & Aucote, 2007). Following the recommendation, item 1 (imaginative) was determined as having the highest modification index. In addition, the factor loadings of item 1 in females and males yielded the highest difference (as shown in Table 3). Vandenberg (2002) stated, "after accurately identifying the items that are not invariant, the researcher engages in a partial metric invariance strategy whereby the non-invariant items are freely estimated in each group, but the invariant items are fixed equal between groups" (p. 151). In light of this suggestion, item 1 was freely estimated in both groups, and then still, a statistically significant difference between this model and configural model ($p < .001$) was observed. The ΔCFI (-.03) value also indicated that the model fit could not be established. Ongoing examination of the item with the highest modification index in the last model was determined as item 6 (inquisitive). In addition, the factor loadings of item 6 in females and males yielded the second-highest difference (as shown in Table 3). When item 1 and item 6 were freely estimated in both groups, an insignificant

difference between this model and configural model (p = .012) at .01 level was found. The ΔCFI value (0.0) lower than -.01 also indicated that the model fit was supported. That is, partial metric invariance was established across the groups, except for the factor loadings of item 1 and item 6.

Table 4. Fit Indexes for Measurement Invariance Models across Gender

Model	χ^2	*df*	CFI	NNFI	RMSEA	$\Delta\chi^2$	Δdf	*p*	ΔCFI
Configural	5426.42	820	.93	.92	.08	-	-	-	-
Metric	7247.01	850	.90	.90	.13	1820.59	30	.000	**-.03**
Partial Metric – I1	5704.39	827	.92	.91	.09	277.97	7	.000	-.01
Partial Metric – I1 & I6	5442.77	826	.93	.93	.09	16.35	6	.012	0.0
Scalar	5456.86	831	.93	.93	.09	14.09	5	.015	0.0
Strict	7286.46	802	.89	.89	.14	1829.6	29	.000	**-.04**

After partial metric invariance was established, the scalar invariance test was conducted. The findings were indicated that the chi-square difference between the scalar model and the partial metric model was not statistically significant ($\Delta\chi^2_{(5)} = 14.09$, $p > .01$). The zero ΔCFI value higher than -.01 indicated scalar invariance. After achieving scalar invariance, in order to examine the highest level of measurement invariance with the test of invariance of error variance was carried on. The chi-square difference between the strict model and the scalar model was statistically significant ($\Delta\chi^2_{(29)} = 1829.6$, $p < .001$) and the ΔCFI = -.04 is lower than -.01. These findings showed that strict invariance was not achieved.

DISCUSSION and CONCLUSION

The aim of the study was twofold. The first purpose of the present study was to test the factorial validity of the Quick Big Five on the Turkish early adulthood sample, and the second was to examine measurement invariance across gender. Firstly, CFA was performed for the whole sample. Afterwards, the model fit was evaluated separately for both male and female groups. Secondly, sequential multiple group CFA tests to examine measurement invariance were conducted.

In general, most of the fit indexes emerged that the Quick Big Five showed adequate fit to the data for the whole sample and the gender groups. However, RMSEA and χ^2/df indicated model misfits. Since the chi-square statistic is sensitive to model size (e.g., the number of observed variables and factors estimated, model degrees of freedom) and sample size (Putnick & Bornstein, 2016), it is not surprising that chi-square showed model misfit. These findings are in line with the findings related to personality traits in the literature. For instance, Beauducel and Wittmann (2005) examined the performance of CFA fit indexes in their simulation study. The simulated data in their study were set as characteristic of data in personality research. As a result of their research, the researchers stated that "there is a tendency to indicate misfit for RMSEA and χ^2/df values when the incremental fit indexes indicate fit." (p.57). They also revealed the situation regarding model fit in personality research as follows:

> According to Raykov (1998), a perfect model fit is not very realistic in personality research because the personality phenomenon can be considered exceedingly complex and because it is not possible to include all relevant variables in studies on personality. When the models do not contain all relevant variables, it is very unlikely that they will explain all relevant aspects of an empirical covariance matrix. Thus, a problem that is emphasized when the application of CFA to personality research is discussed is the extreme complexity of the phenomena under investigation. (Beauducel & Wittmann, 2005, p.42).

As researchers pointed out, it is obvious that there are some problems in model-data fit concerning personality research. The current research findings also are consistent with the literature.

Based on the findings, full configural, partial metric, and scalar invariance were achieved across gender. The fact that configural invariance has been achieved indicates that the Quick Big Five Scale

has a comparable factor structure between females and males. Configural invariance is a prerequisite and should be established in order for subsequent tests to be consequential (Vandenberg & Lance, 2000). In the subsequent test, findings failed to support full metric invariance. However, if latent constructs are to be meaningful in a comparison between groups, equal factor loadings must first be obtained (Cheung & Rensvold, 2002). Therefore, after investigating modification indices, the two items found as non-invariant across the groups. Model fit was acceptable after freeing the factor loadings for item 1 and item 6. The two non-invariant items were "imaginative" and "inquisitive". Both of the items were under the same dimension entitled Openness. Males had higher factor loadings on both non-invariant items which implies that these items are more strongly associated with the scale of the Quick Big Five in males than in females. In other words, these two statements have a different meaning and/or interpretation for the males and the females. This finding is understandable given the patriarchal cultural context of Turkey, and individuals are at the onset of their lives meticulously socialized into highly rigid gender roles where males are encouraged to explore their environments and be independent while female behaviors are closely controlled and monitored so as to promote a strictly rule-abiding lifestyle. Therefore, boys are encouraged and praised for their curiosity and bravery in an exploration of their environment and accumulation of life skills while girls are particularly in the name of "sexual protection" are discouraged toward such exploration whether that be actual or imaginary. In short, males and females are given extremely different sets of rules regarding experimentation with new experiences.

After partial metric invariance was fulfilled, the scalar invariance was tested. The findings showed that item intercepts (except for item 1 and item 6) were invariant across the gender groups. These findings are partly consistent with the findings of the study conducted by Morizot (2014) on an adolescent sample. Morizot (2014) reported that partial scalar (intercept) invariance was achieved when four items were released in the Big Five Personality Trait Short Questionnaire (BFPTSQ). Two of these non-invariance items were artistic-related items that were from the Openness. As mentioned above, in the present study two items of Openness caused metric non-invariance. In accordance with the current literature, the items on Openness had the lowest fit for the FFM data (Rollock & Lui, 2016). There appear some difficulties in understanding the concept of Openness (McCrae & Costa, 1997). Openness is quite hard to define clearly (DeYoung, Peterson, & Higgins, 2005). This may be due to the fact that the abstract and complex definition of Openness (Connelly, Ones, Davies, & Birkland, 2014). Openness includes motivation, needs to reach out novel and varied experience, but sometimes proposes clearly improper receptivity (McCrae & Costa, 1997). Openness to Experience also requires vision, aesthetic sensitivity, and is willing to discard the thought of traditional values. Thus, the dimension of Openness is perhaps not a core concept of personality universally but may have specific meanings in cultural contexts. So much so that the Openness factor did not emerge in the original Chinese Personality Assessment Inventory (Cheung et al., 2008). This was because the FFM model, which was built on the conceptualization of Western-centered personality, did not fit into the more collective Eastern culture (Cheung, Fan, & To, 2008). Triandis and Suh (2002) stated, "The Openness factor is problematic in several studies" and added "Openness emerges more readily in individualist cultures, particularly among student samples that tend to be idiocentric, than in collectivist cultures" (p. 150). There are also views that culture has different levels of influence, even in a single psychological domain such as personality (McAdams & Pals, 2006). McCrae, Yik, Trapnell, Bond, and Paulhus (1998) stated that the cross-lingual equivalence of the scale of Openness was quite limited but still this result was not amazing because it measures the "attitudinal reflections" of the relevant areas of the scale "and attitudes are undoubtedly influenced by the cultural context" (p. 1052).

The highest level of measurement invariance is strict invariance. In the current study the strict invariance was tested but not achieved. This finding was in line with the study done by Samuel et al. (2015) in which they demonstrated full configural, metric, and scalar invariance but did not achieve strict invariance on The Five-Factor Model Rating Form across gender. On the other hand, in the literature, it is noted that strict invariance is a very restricted test; thus, it is not compulsory to compare latent mean differences (Brown, 2006).

In conclusion, the findings of the CFA confirmed the Quick Big Five (five-factor) adequately fit the data from the Turkish early adulthood sample. In addition, each of the 30 items of the scale was

embedded into a related latent factor in both gender groups. This study resulted in several important outcomes. The first important outcome of this study is that the QBF scale operates in Turkish early adulthood sample. Further, the QBF scale was able to carry on full configural, partial metric and scalar invariance between males and females. That is, the QBF scores have the same measurement unit and origin across gender groups when the item 1 and item 6 are excluded. Therefore, the equivalence evidence of the QBF scale of a Turkish sample was built on across gender groups. In other words, meaningful comparisons can be made between the latent mean of the construct.

Even within a nation itself, differences in response manner or expression of personality traits can be shaped depending on cultural contexts (Rollock & Lui, 2016). Therefore, in future research, evidence of validity for diverse groups can be investigated. Likewise, the measurement invariance of the distinct comparison groups can be examined. Because, while examining personality traits, it provides more insight into similarities and differences in item-based studies rather than domains or factors. In addition, there is a need for comprehensive studies on whether the Openness dimension and the facets under this dimension are an etic (universally) or an emic (culture-based) construct. Besides, inconsistency was observed between the CFA fit indices in this study. Therefore, further research on the behavior of different fit indices could be conducted in personality research.

Because personality traits are closely associated with academic variables, educators who intend to enhance individuals' academic performance should have a keen interest in personality. The findings of this study indicated that the QBF is a valid self-report tool that can be easily applied for the early adulthood period in Turkish culture. Thus, the QBF can be used to enhance academic achievement as well as tailoring of teaching methods and techniques to the individual in school settings. Likewise, it can be used at least in addition to other instruments in employee selection in a variety of human resources and occupational guidance settings. In addition, the QBF scores can guide educational and vocational counselors to provide more functional guidance for clients. This research includes some theoretical implications. It confirmed that making group comparisons without taking into account the items where measurement invariance cannot be achieved would lead to biased decisions. It also added new validity evidence to existing personality literature.

REFERENCES

Arnett, J. J. (2000). Emerging adulthood: A theory of development from the late teens through the twenties. *American Psychologist, 55*(5), 469-480. doi: 10.1037//0003-066X.55.5.469

Barrick, M. R., & Mount, M. K. (1991). The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44*(1), 1-26. Retrieved from http://jwalkonline.org/docs/Grad%20Classes/Fall%2007/Org%20Psy/big%205%20and%20job%20perf.pdf

Beauducel, A., & Wittmann, W. W. (2005). Simulation study on fit indexes in CFA based on data with slightly distorted simple structure. *Structural Equation Modeling, 12*(1), 41-75. doi: 10.1207/s15328007sem1201_3

Berings, D., De Fruyt, F., & Bouwen, R. (2004). Work values and personality traits as predictors of enterprising and social vocational interests. *Personality and Individual Differences, 36*(2), 349-364. doi: 10.1016/s0191-8869(03)00101-6

Bidjerano, T., & Dai, D. Y. (2007). The relationship between the big-five model of personality and self-regulated learning strategies. *Learning and individual differences, 17*(1), 69-81. doi: 10.1016/j.lindif.2007.02.001

Bleidorn, W., Hopwood, C. J., & Lucas, R. E. (2018). Life events and personality trait change. *Journal of Personality, 86*(1), 83-96. doi: 10.1111/jopy.12286

Bono, J. E., Boles, T. L., Judge, T. A., & Lauver, K. J. (2002). The role of personality in task and relationship conflict. *Journal of Personality, 70*(3), 311-344. doi: 10.1111/1467-6494.05007

Borghuis, J., Denissen, J. J., Oberski, D., Sijtsma, K., Meeus, W. H., Branje, S., …, & Bleidorn, W. (2017). Big Five personality stability, change, and codevelopment across adolescence and early adulthood. *Journal of Personality and Social Psychology, 113*(4), 641-657. doi: 10.1037/pspp0000138

Borkenau, P., McCrae, R. R., & Terracciano, A. (2013). Do men vary more than women in personality? A study in 51 cultures. *Journal of Research in Personality, 47*(2), 135-144. doi: 10.1016/j.jrp.2012.12.001

Bouchard, T. J., & Loehlin, J. C. (2001). Genes, evolution, and personality. *Behavior Genetics, 31*(3), 243-273. doi: 10.1023/A:1012294324713

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: Guilford Press.

Browne, K. R. (2006). Evolved sex differences and occupational segregation. *Journal of Organizational Behavior, 27*(2), 143-162. doi: 10.1002/job.349

Buss, D. M. (2004). *Evolutionary psychology: The new science of the mind* (2nd ed.). Boston, MA: Allyn & Bacon.

Chamorro-Premuzic, T., & Furnham, A. (2003). Personality predicts academic performance: Evidence from two longitudinal university samples. *Journal of Research in Personality, 37*(4), 319-338. doi: 10.1016/S0092-6566(02)00578-0

Cheung, F. M., Cheung, S. F., Zhang, J., Leung, K., Leong, F., & Huiyeh, K. (2008). Relevance of openness as a personality dimension in Chinese culture: Aspects of its cultural relevance. *Journal of Cross-Cultural Psychology, 39*(1), 81-108. doi: 10.1177/0022022107311968

Cheung, F., Fan, W., & To, C. (2008). The Chinese Personality Assessment Inventory as a culturally relevant personality measure in applied settings. *Social and Personality Psychology Compass, 2*(1), 74-89. doi: 10.1111/j.1751-9004.2007.00045.x

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*(2), 233-255. doi: 10.1207/S15328007SEM0902_5

Cole, D., & Espinoza, A. (2008). Examining the academic success of Latino students in science technology engineering and mathematics (STEM) majors. *Journal of College Student Development, 49*(4), 285-300. doi: 10.1353/csd.0.0018

Connelly, B. S., Ones, D. S., Davies, S. E., & Birkland, A. (2014). Opening up openness: A theoretical sort following critical incidents methodology and a meta-analytic investigation of the trait family measures. *Journal of Personality Assessment, 96*(1), 17-28. doi: 10.1080/00223891.2013.809355

Cooper, A., Gomez, R., & Aucote, H. (2007). The behavioural inhibition system and behavioural approach system (BIS/BAS) scales: Measurement and structural invariance across adults and adolescents. *Personality and Individual Differences, 43*(2), 295-305. doi: 10.1016/j.paid.2006.11.023

Costa Jr, P. T., Terracciano, A., & McCrae, R. R. (2001). Gender differences in personality traits across cultures: robust and surprising findings. *Journal of Personality and Social Psychology, 81*(2), 322-331. doi: 10.1037/0022-3514.81.2.322

De Raad, B., & Schouwenburg, H. C. (1996). Personality in learning and education: A review. *European Journal of Personality, 10*(5), 303-336. doi: 10.1002/(SICI)1099-0984(199612)10:5<303::AID-PER262>3.0.CO;2-2

DeYoung, C. G., Peterson, J. B., & Higgins, D. M. (2005). Sources of openness/intellect: Cognitive and neuropsychological correlates of the fifth factor of personality. *Journal of Personality, 73*(4), 825-858. doi: 10.1111/j.1467-6494.2005.00330.x

Dimitrov, D.M. (2010). Testing for factorial invariance in the context of construct validation. *Measurement and Evaluation in Counseling and Development, 43*(2) 121-149. doi: 10.1177/0748175610373459

Donnellan, M.B., Conger, R.D., & Burzette, R.G. (2007). Personality development from late adolescence to young adulthood: Differential stability, normative maturity, and evidence for the maturity-stability hypothesis. *Journal of Personality, 75*(2), 237-263. doi: 10.1111/j.1467-6494.2007.00438.x

Duckworth, A. L., & Seligman, M. E. P. (2005). Self-Discipline outdoes iq in predicting academic performance of adolescents. *Psychological Science, 16*(12), 939-944. Retrieved from https://www.jstor.org/stable/pdf/40064361.pdf

Dumfart, B., & Neubauer, A. C. (2016). Conscientiousness is the most powerful noncognitive predictor of school achievement in adolescents. *Journal of Individual Differences, 37*(1), 8-15. doi: 10.1027/1614-0001/a000182

Durbin, C. E., Hicks, B. M., Blonigen, D. M., Johnson, W., Iacono, W. G., & McGue, M. (2016). Personality trait change across late childhood to young adulthood: Evidence for nonlinearity and sex differences in change. *European Journal of Personality, 30*(1), 31-44. doi: 10.1002/per.2013

Eccles, J. (2011). Gendered educational and occupational choices: Applying the Eccles et al. model of achievement-related choices. *International Journal of Behavioral Development, 35*(3), 195-201. doi: 10.1177/0165025411398185

Fadjukoff, P., Feldt, T., Kokko, K., & Pulkkinen, L. (2019). Identity status change within personal style clusters: a longitudinal perspective from early adulthood to midlife. *Identity, 19*(1), 1-17. doi: 10.1080/15283488.2019.1566066

Falk, A., & Hermle, J. (2018). Relationship of gender differences in preferences to economic development and gender equality (IZA Discussion Papers, No. 12059). Institute of Labor Economics (IZA), Bonn. Retrieved from http://hdl.handle.net/10419/193353

Figueredo, A. J., Sefcek, J. A., & Jones, D. N. (2006). The ideal romantic partner personality. *Personality and Individual Differences, 41*(3), 431-441. doi: 10.1016/j.paid.2006.02.004

Fischer, A. H., & Manstead, A. S. R. (2000). The relation between gender and emotion in different cultures. In A. H. Fischer (Ed.), *Studies in emotion and social interaction. Second series. Gender and emotion: Social psychological perspectives* (pp. 71-94). Cambridge University Press. doi: 10.1017/CBO9780511628191.005

Funder, D. C. (2001). Personality. *Annual Review of Psychology, 52*, 197-221. Retrieved from https://intranet.newriver.edu/images/stories/library/Stennett_Psychology_Articles/Personality.pdf

Gasser, C. E., Larson, L. M., & Borgen, F.H. (2007). Concurrent validity of the 2005 Strong Interest Inventory: An examination of gender and major field of study. *Journal of Career Assessment, 15*(1), 23-43. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.908.2140&rep=rep1&type=pdf

Guimond, S. (2008). Psychological similarities and differences between women and men across cultures. *Social and Personality Psychology Compass, 2*(1), 494-510. doi: 10.1111/j.1751-9004.2007.00036.x

Heckman, J. J., Stixrud, J., & Urzua, S. (2006). The effects of cognitive and non-cognitive abilities on labor market outcomes and social behavior. *Journal of Labor Economics, 24*(3), 411-482. Retrieved from https://www.nber.org/papers/w12006.pdf

Hu, L. T., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods, 3*(4), 424-453. Retrieved from https://pdfs.semanticscholar.org/a92c/9726361d9c1d165dbf2ea781b6c48363a816.pdfRelated

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1-55. doi: 10.1080/10705519909540118

John, O. P., & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (2nd ed., pp. 102-138). New York, NY: Guilford Press.

John, O. P., Neumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative Big-Five trait taxonomy: History, measurement, and conceptual issues. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of personality: Theory and research* (3rd ed., pp. 114-158). New York, NY: Guilford.

Johnson, J. G., Cohen, P., Brown, J., Smailes, E. M., & Bernstein, D. P. (1999). Childhood maltreatment increases risk for personality disorders during early adulthood. *Archives of General Psychiatry, 56*(7), 600-606. doi: 10.1001/archpsyc.56.7.600

Josephs, R. A., Markus, H. R., & Tafarodi, R. W. (1992). Gender and self-esteem. *Journal of Personality and Social Psychology, 63*(3), 391-402. Retrieved from https://www.researchgate.net/profile/Robert_Josephs/publication/21751337_Gender_and_Self-Esteem/links/0912f5098177bd1614000000.pdf

Judge, T. A., & Ilies, R. (2002). Relationship of personality to performance motivation: A meta-analytic review. *Journal of applied psychology, 87*(4), 797-807. doi: 10.1037//0021-9010.87.4.797

Judge, T. A., Erez, A., Bono, J. E., & Thoresen, C. J. (2002). Are measures of self-esteem, neuroticism, locus of control, and generalized self-efficacy indicators of a common core construct? *Journal of Personality and Social Psychology, 83*(3), 693-710. doi: 10.1037//0022-3514.83.3.693

Kajonius, P. J., & Johnson, J. (2018). Sex differences in 30 facets of the five-factor model of personality in the large public (N=320,128). *Personality and Individual Differences, 129*(July), 126-130. doi: 10.1016/j.paid.2018.03.026

Klein, H. J., & Lee, S. (2006). The effects of personality on learning: The mediating role of goal setting. *Human Performance, 19*(1), 43-66. doi: 10.1207/s15327043hup1901_3

Klimstra, T. A., Luyckx, K., Branje, S., Teppers, E., Goossens, L., & Meeus, W. H. (2013). Personality traits, interpersonal identity, and relationship stability: Longitudinal linkages in late adolescence and young adulthood. *Journal of Youth and Adolescence, 42*(11), 1661-1673. doi: 10.1007/s10964-012-9862-8

Kline, R. B. (2011). *Principles and practice of structural equation modeling.* New York, NY: The Guilford Press.

Korkmaz, M., Somer, O., & Gungor, D. (2013). Measurement equivalence across gender with mean and covariance structure of five factor personality inventory for adolescent sample. *Education and Science, 38*(170). 121-134. Retrieved from http://egitimvebilim.ted.org.tr/index.php/EB/article/view/1532/550

Langen, A. V., & Dekkers, H. (2005). Cross-national differences in participating in tertiary science, technology, engineering and mathematics education. *Comparative Education, 41*(3), 329-350. doi: 10.1080/03050060500211708

Legewie, J., & DiPrete, T. A. (2014). The high school environment and the gender gap in science and engineering. *Sociology of Education, 87*(4), 259-280. doi: 10.1177/0038040714547770

Lippa, R. A. (2010). Gender differences in personality and interests: When, where, and why? *Social and Personality Psychology Compass, 4*(11), 1098-1110. doi: 10.1111/j.1751-9004.2010.00320.x

Lounsbury, J. W., Sundstrom, E., Loveland, J. M., & Gibson, L. W. (2003). Intelligence, "Big Five" personality traits, and work drive as predictors of course grade. *Personality and Individual Differences, 35*(6), 1231-1239. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.936.3275&rep=rep1&type=pdf

Low, K., & Rounds, J. (2007). Interest change and continuity from early adolescence to middle adulthood. *International Journal for Educational and Vocational Guidance, 7*(1), 23-36. doi: 10.1007/s10775-006-9110-4

Lynn, R. (1993). Sex differences in competitiveness and the valuation of money in twenty countries. *Journal of Social Psychology, 133*(4), 507-511.

Manders, W. A., Scholte, R. H., Janssens, J. M., & De Bruyn, E. E. (2006). Adolescent personality, problem behaviour and the quality of the parent–adolescent relationship. *European Journal of Personality, 20*(3), 237-254. doi: 10.1002/per.574

Matthews, G., Zeidner, M., & Roberts, R. D. (2006). Models of personality and affect for education: A review and synthesis. In P. A. Alexander & P. H. Winne (Eds.), *Handbook of educational psychology* (2nd ed., pp. 163-186). Mahwah, NJ: Erlbaum.

McAdams, D. P., & Pals, J. L. (2006). A new Big Five: fundamental principles for an integrative science of personality. *American Psychologist, 61*(3), 204-217. Retrieved from http://people.wku.edu/richard.miller/new%20big%20five.pdf

McCrae, R. R., & Costa, P. T., Jr. (1997). Conceptions and correlates of openness to experience. In R. Hogan, J. A. Johnson, & S. R. Briggs (Eds.), *Handbook of personality psychology* (pp. 825-847). Academic Press. doi: 10.1016/B978-012134645-4/50032-9

McCrae, R. R., Terracciano, A., & Pro, P. P. C. (2005). Personality profiles of cultures: Aggregate personality traits. *Journal of Personality and Social Psychology, 89*(3), 407-425. doi: 10.1037/0022-3514.89.3.407

McCrae, R. R., Yik, M. S., Trapnell, P. D., Bond, M. H., & Paulhus, D. L. (1998). Interpreting personality profiles across cultures: Bilingual, acculturation, and peer rating studies of Chinese undergraduates. *Journal of Personality and Social Psychology, 74*(4), 1041-1055. doi: 10.1037/0022-3514.74.4.1041

Milfont, T. L., & Fischer, R. (2010). Testing measurement invariance across groups: Applications in cross-cultural research. *International Journal of psychological research, 3*(1), 111-130. Retrieved from http://45.5.172.45/bitstream/10819/6503/1/Testing_Measurement_Invariance_Milfont_2010.pdf

Morizot, J. (2014). Construct validity of adolescents' self-reported big five personality traits: Importance of conceptual breadth and initial validation of a short measure. *Assessment, 21*(5), 580-606. doi: 10.1177/1073191114524015

Morsunbul, U. (2014). The validity and reliability study of the Turkish version of quick big five personality test. *Dusunen Adam The Journal of Psychiatry and Neurological Sciences, 27*(4), 316-322. doi: 10.5350/DAJPN2014270405

Mueller, G., & Plug, E. (2006). Estimating the effect of personality on male and female earnings. *Industrial and Labor Relations Review, 60*(1), 3-22. Retrieved from http://www.jstor.org/stable/25067572

Nguyen, N. T., Allen, L. C., & Fraccastoro, K. (2005). Personality predicts academic performance: Exploring the moderating role of gender. *Journal of Higher Education Policy and Management, 27*(1), 105-117. doi: 10.1080/13600800500046313

Noftle, E. E., & Robins, R. W. (2007). Personality predictors of academic outcomes: Big five correlates of GPA and SAT scores. *Journal of Personality and Social Psychology, 93*(1), 116-130. doi: 10.1037/0022-3514.93.1.116

O'Connor, M. C., & Paunonen, S. V. (2007). Big five personality predictors of post-secondary academic performance. *Personality and Individual Differences, 43*(5), 971-990. doi: 10.1016/j.paid.2007.03.017

Ock, J., McAbee, S. T., Mulfinger, E., & Oswald, F. L. (2020). The practical effects of measurement invariance: gender invariance in two big five personality measures. *Assessment, 27*(4), 657-674. doi: 10.1177/1073191119885018

Paunonen, S. V., & Ashton, M. C. (2001). Big five factors and facets and the prediction of behavior. *Journal of Personality and Social Psychology, 81*(3), 524-539. doi: 10.1037/0022-3514.81.3.524

Poropat, A. E. (2009). A meta-analysis of the five-factor model of personality and academic performance. *Psychological Bulletin, 135*(2), 322-338. doi: 10.1037/a0014996

Poropat, A. E. (2011). The Eysenckian personality factors and their correlations with academic performance. *British Journal of Educational Psychology, 81*(1), 41-58. doi: 10.1348/000709910X497671

Poropat, A. E. (2014). Other-rated personality and academic performance: Evidence and implications. *Learning and Individual Differences, 34*, 24-32. doi: 10.1016/j.lindif.2014.05.013

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review, 41*, 71-90. doi: 10.1016/j.dr.2016.06.004

Rindermann, H., & Neubauer, A. (2001). The influence of personality on three aspects of cognitive performance: Processing speed, intelligence and school performance. *Personality and Individual Differences, 30*(5), 829-842. Retrieved from https://pdfs.semanticscholar.org/e468/6271719ef3a12ffc1e3db46bb705cf2195a6.pdf

Roberts, B. W., Kuncel, N., Shiner, R. N., Caspi, A., & Goldberg, L. R. (2007). The power of personality: The comparative validity of personality traits, socio-economic status, and cognitive ability for predicting important life outcomes. *Perspectives in Psychological Science, 2*(4), 313-345. doi: 10.1111/j.1745-6916.2007.00047.x

Roberts, B. W., Walton, K. E., & Viechtbauer, W. (2006). Patterns of mean-level change in personality traits across the life course: a meta-analysis of longitudinal studies. *Psychological Bulletin, 132*(1), 1-25. doi: 10.1037/0033-2909.132.1.1

Roets, A., & Van Hiel, A. (2011). Item selection and validation of a brief, 15-item version of the need for closure scale. *Personality and Individual Differences, 50*(1), 90-94. doi: 10.1016/j.paid.2010.09.004

Rollock, D., & Lui, P. P. (2016). Measurement invariance and the five-factor model of personality: Asian international and Euro American cultural groups. *Assessment, 23*(5), 571-587. doi: 10.1177/1073191115590854

Samuel, D. B., South, S. C., & Griffin, S. A. (2015). Factorial invariance of the five-factor model rating form across gender. *Assessment, 22*(1), 65-75. doi: 10.1177/1073191114536772

Sass, D. A. (2011). Testing measurement invariance and comparing latent factor means within a confirmatory factor analysis framework. *Journal of Psychoeducational Assessment, 29*(4), 347-363. doi: 10.1177/0734282911406661

Saucier, G., & Goldberg, L. R. (1996). The language of personality: Lexical perspectives on the five- factor model. In J. S. Wiggins. (Ed.), *The five-factor model of personality: Theoretical Perspective* (pp. 21-50). New York, NY; Guilford.

Schmitt, D. P., Long, A. E., McPhearson, A., O'Brien, K., Remmert, B., & Shah, S. H. (2017). Personality and gender differences in global perspective. *International Journal of Psychology, 52*(1), 45-56. doi: 10.1002/ijop.12265

Schmitt, D. P., Realo, A., Voracek, M., & Allik, J. (2008). Why can't a man be more like a woman? Sex differences in Big Five personality traits across 55 cultures. *Journal of Personality and Social Psychology, 94*(1), 168-182. doi: 10.1037/0022-3514.94.1.168

Shiner, R. L., Allen, T. A., & Masten, A. S. (2017). Adversity in adolescence predicts personality trait change from childhood to adulthood. *Journal of Research in Personality, 67*, 171-182. doi: 10.1016/j.jrp.2016.10.002

Soto, C. J. (2016). The little six personality dimensions from early childhood to early adulthood: Mean-level age and gender differences in parents' reports. *Journal of Personality, 84*(4), 409-422. doi: 10.1111/jopy.12168

Tabachnick, B.G., & Fidell, L.S. (2001). *Using multivariate statistics* (4th ed.). Needham Heights, MA: Allyn & Bacon.

Triandis, H. C., & Suh, E. M. (2002). Cultural influences on personality. *Annual Review of Psychology, 53*(1), 133-160. doi: 10.1146/annurev.psych.53.100901.135200

Van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology, 9*(4), 486-492. doi: 10.1080/17405629.2012.686740

Vandenberg, R. J. (2002). Toward a further understanding of and improvement in measurement invariance methods and procedures. *Organizational Research Methods, 5*(2), 139-158. doi: 10.1177/1094428102005002001

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational research methods, 3*(1), 4-70. doi: 10.1177/109442810031002

Vermults, A. A., & Gerris, J. R. M. (2005). *QBF: Quick Big Five persoonlijkheidstest handleiding [Quick Big Five personality test manual]*. Leeuwarden, the Netherlands: LDC Publications.

Wang, M. T., & Degol, J. L. (2017). Gender gap in science, technology, engineering, and mathematics (STEM): Current knowledge, implications for practice, policy, and future directions. *Educational Psychology Review, 29*(1), 119-140. doi: 10.1007/s10648-015-9355-x

Willingham, W. W., Pollack, J. M., & Lewis, C. (2002). Grades and test scores: Accounting for observed differences. *Journal of Educational Measurement, 39*(1), 1-37. Retrieved from https://www.jstor.org/stable/1435104?seq=1&cid=pdf-reference#references_tab_contents

Hızlı Büyük Beşli Kişilik Testi: Cinsiyete Göre Ölçme Değişmezliğinin İncelenmesi

Giriş

Kişilik ölçümünde yaygın olarak kullanılan kavramsallaştırma Beş Faktör Modelidir. Bu model kişiliği beş özellik alanına göre organize eder. Araştırmacılar Beş Faktör Kişilik Modeli'nin neredeyse evrensel düzeyde temsiliyeti üzerinde büyük ölçüde uzlaşmaya varmış durumdadır (John, Neumann & Soto, 2008; Korkmaz, Somer & Gungor, 2013; McCrae, Terracciano & Pro, 2005).

Beş Faktör Modeli'nin kuramsal temelleri sözcük (lexical) hipotezi ile oluşturulmuştur. Bu hipoteze göre; insanların kişilik özelliği olarak en çok öne çıkan özellikleri önünde sonunda dillerinin bir parçası olur ve kullandıkları dilde de kendilerini gösterir. Bu hipotezden yola çıkılarak kişilik özelliklerini dillerdeki betimleyici sıfatlara bakarak belirlemek mümkün görülmüştür. Başta İngilizce olmak üzere kişiliğin göstergeleri olabilecek sıfatlar belirlenmiş sonra da başka dillerde faktör analitik çalışmalarla Beş Faktör Modeli'ne dayalı ölçekler geliştirmek ve geçerliğini incelemek mümkün olmuştur (Saucier & Goldberg, 1996). Bunlardan biri de Büyük Beşli'dir. Büyük Beşli boyutları uyumluluk, sorumluluk, duygusal denge, dışadönüklük ve deneyime açıklık olarak belirlenmiştir.

Kişilik gelişimi üzerine yapılan çoğu araştırma erken yetişkinlik dönemine odaklanmıştır. Bunun nedeni, kişilik gelişiminin beyin gelişimine bağlı olarak 25 hatta 28 yaşına kadar devam etmesidir. Bir diğer nedeni de yetişkinliğe geçişteki 18 ile 30 yaş arasının samimiyet, girişimcilik, sosyal ilgiler, kimlik, iş ve ebeveynlik açısından önemli bir gelişim evresi olmasıdır (Arnett, 2000). Araştırmalar erken yetişkinlik döneminde ilgi alanlarının kristalize olduğunu ve dengelediğini ayrıca kariyer hedeflerinin ve ileriye dönük beklentilerinin kişisel ve çevresel özelliklere uyum sağlama açısından daha gerçekçi hale geldiğini göstermiştir (Low & Rounds, 2007).

Kadınlar ve erkekler arasındaki psikolojik farklılıklar her zaman incelenen bir konu olmuştur (Kajonius & Johnson, 2018). Peki, kişilikteki cinsiyet farklılığını incelemek neden önemlidir? Öncelikle kişilik üzerindeki cinsiyet farklılıkları kültürler arası tüm araştırmalarda gözlenmiştir. Bu nedenle evrensel bir husustur. Bir diğeri, kişilikteki cinsiyet farklılıklarının yaşam süresi boyunca istikrar göstermesidir (Donnellan, Conger, & Burzette, 2007). Bu da bize bireylerin gelecekteki seçimlerinin eğilimi ve bu seçimler sonucunda karşı karşıya kalacakları durumlar hakkında bilgi verir. Ayrıca mesleki, eğitsel, eş seçme, çatışma, ilişki düzenleme gibi sosyal pek çok seçimler kişilikle ilişkilidir (Berings, De Fruyt, & Bouwen, 2004; Bono, Boles, Judge, & Lauver, 2002; Figueredo, Sefcek, & Jones, 2006; Gasser, Larson, & Borgen, 2007). Bunun yanında meta-analiz çalışmalar da psikolojik değişkenler üzerindeki cinsiyet farklılıklarının incelenen yapıya göre değişkenlik gösterdiğini ortaya çıkarmaktadır. Böylece, kişilik özellikleri gibi psikolojik etmenlerin incelenmesi yoluyla bireylerin özellikle de kadınların eğitim, beceri ve mesleki açıdan gelişimlerinin izlenmesi ve iyileştirilebilmesi mümkün olabilir.

Kişilik ve kişiliğin sosyal ve ekonomik yapılarla ilişkileri her daim canlı bir araştırma konusu olmuştur (Funder, 2001). Bir işte veya akademik faaliyetlerde performans gösterme isteği ve performansta devamlılık zihinsel yetenekten ziyade (Heckman, Stixrud, & Urzua, 2006, Willingham, Pollack, & Lewis, 2002) kişilik faktörleri tarafından daha belirleyici bulunmuştur (Judge & Ilies, 2002). Literatürde bazı çalışmalar zihinsel olmayan becerilerin çocukların ve ergenlerin okul performanslarında önemli rol oynadığını göstermiştir (Duckworth & Seligman, 2005). Açıkçası, öğrencilerin akademik performanslarını incelemek oldukça önemlidir, çünkü toplumlar ve bireyler tarafından eğitime önemli yatırımlar yapılmakta, bu da eğitim performansına verilen yüksek değeri göstermektedir (Poropat, 2009). Akademik performansla Büyük Beşli kişilik faktörleri arasında güçlü ilişkilerin olması da eğitsel açıdan kişilik özelliklerine daha fazla eğilmemiz gerektiğine işaret etmektedir.

Farklı kültürler ve örneklemlerle yapılan ampirik çalışmalarla Büyük Beşli kişilik faktör yapısının sağlamlığı desteklenmiştir. Ancak bir psikolojik yapının karşılaştırma grupları arasında farklılık veya

benzerlikleri yorumlanmak isteniliyorsa öncelikle ölçme değişmezliği yoluyla psikolojik yapının değişmezliğinin test edilmesi gerekir. Bu nedenle hemen her disiplinde (psikoloji, sağlık, ekonomi, eğitim, sosyoloji vb.) yaygın olarak kullanılan Büyük Beşli faktör yapısının daha ileri geçerlik analizleri ile desteklendiği araştırmalara ihtiyaç vardır. Büyük Beşli için karşılaştırma gruplarında ancak skaler ölçme değişmezliği sağlanabilirse alt gruplardan elde edilen puanlar (veya gizil ortalamalar) arasında anlamlı karşılaştırmalar yapılabilmesi mümkün olur (Ock, McAbee, Mulfinger, & Oswald, 2019; Sass, 2011). Aksi takdirde ortaya çıkan farklılıkların gerçekten gruplar arasındaki farklılığa mı yoksa psikolojik yapının eşdeğer olmayışından kaynaklı bir duruma mı atfedilip atfedilemeyeceği belirlenemez. Bu durumda psikolojik yapının hem geçerliği hem de genellenebilirliği sorunlu hale gelir.

Türkiye'de Büyük Beşli kuramına göre yapılandırılan ölçekler olmasına rağmen sadece Korkmaz ve diğerleri (2013) geliştirdikleri 200 maddelik ölçek üzerinden lisede öğrenim gören ergen gruplarında cinsiyete göre ölçme değişmezliğini incelemişlerdir. Ancak bu konuda daha fazla araştırmalara ihtiyaç vardır. Yaşam dönemleri boyunca kişilik özelliklerinin gelişimini anlamak kuramsal ve pratik sonuçlar içerir (Roberts, Kuncel, Shiner, Caspi, & Goldberg, 2007). Özellikle görece daha kısa kendini rapor etme araçlarından elde edilen puanların ileri çalışmalarla geçerliğinin incelenmesi önem arz etmektedir. Bu nedenle, bu çalışmada pek çok disiplinde kullanılan Hızlı Büyük Beşli faktörlerinin Türk erken yetişkin örnekleminde geçerliği ve cinsiyete göre ölçme değişmezliğinin test edilmesi amaçlanmıştır. İkinci amaç doğrultusunda Hızlı Büyük Beşli faktörlerinin cinsiyete göre yapısal, metrik, skaler ve katı ölçme değişmezliği araştırılmıştır.

Yöntem

Bu araştırmaya yaşları 17-32 arasında değişen İç Anadolu Bölgesi'nde öğrenim gören 1114 üniversite öğrencisi katılmıştır. Katılımcıların 659'ü kadın (%59) ve 455'i erkek (%41) olduğunu beyan etmiştir. Kişilik özelliklerini ölçmek için Vermulst ve Gerris (2005) tarafından geliştirilen Hızlı Büyük Beşli Kişilik ölçeği kullanılmıştır. Ölçek 30 maddeden oluşmaktadır. Her kişilik özelliği altı maddeyle ölçülmektedir. Alt ölçekler için Cronbach Alfa değerleri .73 ile .88 arasında değişmektedir. Ölçek Morsunbul (2014) tarafından Türk kültürüne uyarlanmıştır. Uyarlama çalışmasında alt ölçeklerin Cronbach Alfa katsayıları .71 ile .81 arasında değiştiği rapor edilmiştir. Doğrulayıcı faktör analizi (DFA) ve çok gruplu DFA analizleri LISREL9.2 programı ile gerçekleştirilmiştir.

Bu çalışmada ölçme değişmezliği çoklu grup doğrulayıcı faktör analiziyle test edilmiştir. Ölçme değişmezliğinin test edilmesinde aşamalı olarak devam eden süreçler vardır. İlk aşamada karşılaştırma grupları için ayrı ayrı DFA yapılarak ölçme modeli test edilir. Eğer model uyumu sağlanırsa, ikinci aşamada söz konusu gruplar için yapısal değişmezlik, metrik değişmezlik, faktör kovaryansları (skaler) değişmezliği ve hata varyansları (katı) değişmezliği sınanır (Dimitrov, 2010). Her bir model, bir önceki model ile karşılaştırılır. Bu iç içe yuvalanmış modelleri karşılaştırmak için ki-kare fark testi kullanılır (Brown, 2006; Dimitrov, 2010; Tabachnick & Fidell, 2001). Her bir model için manidar bir farkın olmaması, ölçme değişmezliğin sağlandığını gösterir.

Ki-kare testinin örneklem büyüklüğüne duyarlı olması nedeniyle iç içe yuvalanmış model karşılaştırmalarında daha dirençli bir gösterge olan CFI fark değerlerinin kullanılması önerilmektedir (Cheung & Rensvold, 2002). Ölçme değişmezliğin sağlanamadığı durumlarda kısmi değişmezlik incelenmelidir. Kısmi değişmezlik sürecinde en büyük modifikasyon üreten parametreler belirlenir. Bu parametreler tek tek serbest bırakılarak değişmezliğin sağlanıp sağlanmadığı incelenir.

Sonuç ve Tartışma

Genel olarak, uyum indekslerinin çoğu, Hızlı Büyük Beşli'nin tüm örneklem ve cinsiyet grupları için verilere yeterli uyum gösterdiğini ortaya çıkarmıştır. Ancak, RMSEA ve χ^2/sd değerlerinde bir miktar model uyumsuzluğu gözlenmiştir. Ki-kare istatistiğinin model büyüklüğüne ve örneklem büyüklüğüne duyarlı olmasından dolayı (Putnick & Bornstein, 2016), ki-kareye bağlı değerlerde model uyumsuzluğunun izlenmesi şaşırtıcı değildir. Bu bulgular, literatürdeki kişilik özelliklerine ilişkin

bulgular ile uyumludur. Örneğin, Beauducel ve Wittmann (2005) DFA uyum indekslerinin simülasyon çalışmalarındaki performansını incelemişlerdir. Araştırmaları sonucunda araştırmacılar RMSEA ve χ^2/sd değerleri için uyumsuzluk gösterme eğilimi olduğunu belirtmişlerdir.

Bulgular, cinsiyete göre tam yapısal, kısmi metrik ve skaler değişmezlik sağlandığını göstermiştir. Yapısal değişmezliğin sağlanmış olması Hızlı Büyük Beş ölçeğinin kadın ve erkekler arasında karşılaştırılabilir faktör yapısına sahip olduğunu belirtir. Bir sonraki aşamada, tam metrik değişmezliği incelenmiştir. Ancak tam metrik değişmezliğin sağlanmadığı ortaya çıkmıştır. Bu nedenle, kısmı değişmezlik incelenmiştir. En büyük modifikasyon indeksi üreten madde deneyime açıklık faktörü altındaki "hayal gücü geniş" maddesi olarak belirlenmiştir. Bu maddeye ilişkin faktör yükleri serbest bırakılarak tekrar metrik değişmezlik incelendiğinde yine değişmezliğin sağlanamadığı görülmüştür. Devam eden süreçte en büyük modifikasyon indeksi üreten bir sonraki madde olan "meraklı" maddesinin faktör yükleri gruplar arasında serbest bırakılmıştır. Çoklu grup DFA bulguları, bu iki madde serbest bırakıldığında kısmi metrik değişmezliğin sağlandığını göstermiştir. Bu iki maddeye ilişkin parametreler serbest bırakıldığında skaler değişmezliğin de sağlandığı gözlenmiştir.

Cinsiyet grupları arasında faktör yük değerleri incelendiğinde erkekler Deneyime Açıklık boyutundaki her iki madde üzerinde de ("hayal gücü geniş" ve "meraklı") daha yüksek değerler elde etmişlerdir. Bu bulgu, erkeklerde söz konusu bu iki maddenin gizil yapı ile daha güçlü bir şekilde ilişkili olduğunu ifade etmektedir. Başka bir deyişle, bu iki maddenin erkekler ve kadınlar için farklı bir anlamı mevcuttur. Türkiye'nin ataerkil kültürel bağlamı göz önüne alındığında bu bulgu anlaşılabilirdir. Nitekim, bu toplumda erkekler doğdukları andan itibaren çevrelerini keşfetmeye ve bağımsız olmaya teşvik edilirken bilakis kadınların davranışları yakından kontrol edilip sürekli takip edilmektedir. Cinsiyet rolleri kadınlar için kurallara uyan bir yaşam tarzını sosyal hayatlarına işlemektedir. Bu nedenle, kızlar, özellikle "cinsel koruma" adı altında, çevrenin keşfi ve yeni yaşam becerileri elde etme fırsatlarını değerlendirme yönünde sürekli bir engelleme ile karşılaşırken, erkelerin yeni deneyimler konusundaki merakları cesaretlendirilir ve övülürler. Kısacası, erkeklere ve kadınlara yeni deneyimler elde etme konusunda son derece farklı kurallar verilir. Bu nedenle deneyime açıklık boyutundaki bu iki maddenin cinsiyet gruplarında eşdeğer anlamları karşılamıyor oluşu anlaşılırdır.

Kısmi metrik değişmezlik sağlandıktan sonra skaler değişmezlik test edilmiştir. Bulgular, açıklık faktörü altındaki iki madde hariç diğer maddelerin, cinsiyet grupları arasında değişmez olduğunu göstermiştir. Bu bulgular, bir ergen örneği üzerinde Morizot (2014) tarafından yapılan çalışmanın bulgularıyla kısmen uyumludur. Morizot (2014) Büyük Beşli Kişilik Özellik Kısa Anketi'nde dört madde serbest bırakıldığında kısmi skaler değişmezliğin sağlandığını bildirmiştir. Bu dört maddeden ikisi, Açıklık boyutuyla ilişkiliydi ve metrik değişmezliğin sağlanamamasına neden olmuştu. Mevcut literatüre göre, Açıklık ile ilgili maddeler Büyük Beşli Modeli'nde en düşük uyuma sahip olarak ortaya çıkmaktadır (Rollock & Lui, 2016). Öyle ki, özgün Çin Kişilik Değerlendirme Envanterinde Açıklık faktörü hiç ortaya çıkmamıştır (Cheung ve diğerleri, 2008). Bunun nedeni, Batı merkezli kişiliğin kavramsallaştırılması üzerine inşa edilen Büyük Beşli Modeli'nin daha kolektif Doğu kültürüne uymaması olarak belirtilmiştir (Cheung, Fan & To, 2008). Triandis ve Suh (2002) Açıklık faktörünün, bireysel kültürlerde, daha kolay ortaya çıktığını belirtmişlerdir. Ayrıca, kişilik gibi tek bir psikolojik alanda bile kültürün farklı düzeylerde etkiye sahip olduğuna dair görüşler vardır (McAdams & Pals, 2006). McCrae, Yik, Trapnell, Bond ve Paulhus (1998), Açıklık boyutunun çapraz dil eşdeğerliğinin oldukça sınırlı olduğunu, ancak bu sonucun şaşırtıcı olmadığını, çünkü ölçeğin ilgili alanlarının tutumsal yansımalarını ölçtüğü ve de tutumların kuşkusuz kültürel bağlamdan etkilendiğini belirtmişlerdir.

En üst düzeydeki ölçme değişmezliği katı değişmezliktir. Bu çalışmada katı değişmezlik test edilmiş ancak sağlanamamıştır. Literatürde katı değişmezliğin çok kısıtlı bir test olduğu belirtilmektedir, bu nedenle gruplar arasında gizil ortalamalar karşılaştırılırken katı değişmezliğin sağlanması zorunlu değildir (Brown, 2006).

Bu çalışma önemli sonuçlar içermektedir. İlk olarak, DFA bulguları, Hızlı Büyük Beşli'nin hem tüm örneklemde hem de kadın ve erkek katılımcılar için model veri uyumunun doğruladığını göstermiştir. Bu çalışmanın ikinci önemli sonucu, Hızlı Büyük Beşli ölçeğinin Türk erken yetişkin örnekleminde işlev gösterdiğinin ortaya konmasıdır. Ayrıca, Hızlı Büyük Beşli ölçeğinde erkekler ve kadınlar

arasında tam yapısal, kısmi metrik ve skaler değişmezlik elde edilmiştir. Bu sonuç, iki madde dışarıda tutulmak suretiyle cinsiyet grupları arasında gizil değişken ortalamalarına ilişkin anlamlı karşılaştırmaların yapılabileceğini belirtmektedir. Unutmamak gerekir ki, ölçme değişmezliğinin sağlanamadığı maddeleri dikkate almadan grup karşılaştırmaları yapmak yanlı kararlara yol açabilecektir. Bu çalışma, mevcut kişilik araştırmalarına yeni geçerlik kanıtları eklemiştir. Gelecekteki çalışmalarda, farklı karşılaştırma grupları için geçerlik kanıtı araştırılabilir ve ölçme değişmezliği incelenebilir.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2020; 11(2); 199-218

Four-Skill Assessment of Turkish Language: Results from a Pilot Project

Emine EROĞLU * H. Eren SUNA ** Hande TANBERKAN ***
Amine CANIDEMIR **** Umare ALTUN *****
Mahmut ÖZER ******

Abstract
This study analyses the results of the 'Four-Skill Test in Turkish Language' (FSTTL) project conducted by the Ministry of National Education to assess the language skills of students as a pilot project and investigates the effects of various variables on language skills. Relationships between language scores and school type, gender, preschool participation, parents' level of education, and course grades are investigated in this descriptive study. The sample is consisted of 1932 students in seventh grade who participated in the pilot study. Test battery, consisted of reading, listening, writing, and speaking subtests, is used to assess the language skills of students within the scope of the FSTTL. Findings show that students in imam-hatip middle schools and middle schools performed at a similar level in all subtests. Female students performed significantly higher than male students in all subtests. Students participated in pre-school education performed significantly higher than those who did not participate in reading, writing, and listening subtests. Findings also show that the increase in parents' level of education leads to an increase in students' subtest scores. The effect of parents' level of education on subtest scores is comparatively higher than the effects of other factors in focus. Significant correlations have been obtained between the four-skill scores and student's Turkish course, social sciences, mathematics, and science course grades. It is suggested that FSTTL must be developed based on the experiences of the pilot project as a standardized test in accordance with the international standards and actively used to improve educational processes.

Keywords: Language Skills, Four-Skill Test in Turkish Language, Language Teaching, Assessment

INTRODUCTION

Language is a living entity that provides communication between people, is dynamic, has its own specific rules, a system of secret treaties that it is not known when it was formed, and a social structure consisting of sounds (Ergin, 1998, p. 2). Language is the basic tool for people to engage with the environment and to express their thoughts and feelings. Any verbal and written reaction of the individual who perceives the events and actions in his environment is directly related to his language skills. In this respect, language skills are among the most basic skills expected for the individual to be able to adapt to daily life, to interact with his environment as an individual and to be a part of social life (Jing, 2006).

* Ministry of National Education, Ankara-Turkey, emineeroglu34@gmail.com, ORCID ID: 0000-0001-6611-3313
** PhD., Ministry of National Education, Ankara- Turkey, herensuna@gmail.com, ORCID ID: 0000-0002-6874-7472
*** PhD., Ministry of National Education, Ankara- Turkey, handetanberkan@gmail.com, ORCID ID: 0000-0001-7142-5397
**** Ministry of National Education, Ankara- Turkey, aminecanidemir@hotmail.com, ORCID ID: 0000-0003-0705-1604
***** Ministry of National Education, Ankara- Turkey, umarealtun@gmail.com, ORCID ID: 0000-0001-8995-0763
****** Prof. PhD., Ministry of National Education, Ankara- Turkey, mahmutozer2002@yahoo.com, ORCID ID: 0000-0001-8722-8670

To cite this article:
Eroğlu, E., Suna, H.E.,Tanberkan, Canıdemir A., Altun, U. & Özer, M.(2020). Four-skill assessment of Turkish language: Results from a pilot project. *Journal of Measurement and Evaluation in Education and Psycholog, 11*(2), 199-218. doi: 10.21031/epod.687758

Geliş Tarihi: 11.02.2020
Kabul Tarihi: 26.03.2020

Individuals use their acquired language skills for social and academic purposes. It is aimed to improve the language skills of individuals both in the social context they will use in daily life and the academic context that they will use throughout their education. Therefore, education systems are structured in such a way that individuals can improve their language skills in both social and academic contexts. Thus, it is aimed to raise individuals who can actively participate in the society, express their feelings and thoughts as they wish, and have literacy skills (Bayyurt, 2013; Cook, 1999). It is also important to note that individuals who have higher levels of language skills also have a significant advantage in employment in diverse sectors (Budria, Colino & Matinez de Ibarreta, 2019; Gazzola & Mazzacani, 2019). Recently, since widespread automation in the labor market is supported by artificial intelligence technologies (Perc, Ozer & Hojnik, 2019), language skills become a much more important factor for adaptability in new circumstances. In this manner, language skills still have a crucial role in communication between people towards the demands of the labor market.

Gaining language skills, which are the basic means for individuals to express their feelings and thoughts, plays an important role in the language to live and to be delivered to the next generations in a proper way. In order for a language to be properly learned, individuals must have gained reading, writing, and listening skills as well as speaking (MoNE, 2019a). Therefore, the acquisition of language skills requires the development of four basic language skills simultaneously (Gautam, 2019; Manaj-Sadiku, 2015). Verbal speech on any subject, texts read to learn, news listened to in daily life, or texts written in order to express their opinions provide individuals to meet their different needs. Four basic language skills, reading, writing, listening and speaking, have a natural relationship with each other, and the development of one skill positively affects the development of other language skills (Brown, 2001; Chengyu, 2018; Gautam, 2019).

Each of the four basic language skills ensures that different functions of the language are performed. Children learn their native language primarily through listening. This learning is also the basis of the individual's ability to learn the native language. Meaning and sounds come to the fore in reading skills. Some symbols need to be analyzed and interpreted to improve reading skills. Writing skill refers to the transformation of emotions, thoughts, opinions, and dreams into text. In writing skill, it is important for the individual to express what they saw, heard, thought, and lived in text. One of the general objectives of the Turkish curriculum was expressed as "to provide students with the ability and habit of describing what they see, watch, listen, read, examine and think, design with words or writing correctly and in accordance with the purpose". The accurate written communication depends on the fulfilment of the external structure, internal structure (narration), spelling, and punctuation dimensions (Deniz, 2000; Kantemir, 1997; Özkırımlı, 1994). Speaking skills can be explained as a set of skills that enable the individual to communicate in the target language (Barın, 1997). Speaking skill is considered as one of the most frequently used language skills of the individual to communicate in a social and academic context (Boonkit, 2010). As can be seen, four basic language skills are considered as components of the language skill of the individual. Models used in the development of language skills and assessment methods of language skills in Turkey is described respectively.

Four Basic Language Skill Approaches and Global Trends in Measuring Language Skill

The development of language skills has been one of the most important issues in education. Many different methods have been developed for the development of language skills, which is a basic communication requirement, and two of these methods are frequently used (Gautam, 2019; Widdowson, 1978). The first of these methods is the behavioural model (major skills model) that divides language skills into subskills such as reading, listening, speaking, and writing and focusing on the development of these subskills separately. In this approach, it is accepted that there is a natural link between basic skills, but each component is developed within itself (Akram & Malik, 2010; Hinkel, 2010). In the integrated model, language skill is seen as a whole with all subskills, and subskills are tried to be developed with the same methods (Xue-Ping, 1997). Both models have their own advantages and limitations, and the approaches used in education systems differ.

Although they are in an organic relationship with each other, the benefits of addressing these skills separately for the development of four basic language skills have been demonstrated by linguists based on data (Hinkel, 2010). Addressing basic skills separately in language teaching enables different methods to be used in developing these skills. In addition, individuals' gains, strengths, and aspects that are open to development can be examined separately according to their language skills (Hinkel, 2002; Stern, 1983). For example, a personalized development plan can be presented to an individual who has sufficient listening and speaking skills, but not sufficient writing and listening skills.

Linguistic scientists express that with the development of language skills separately, students can understand different layers of language faster and use different skills more effectively (Canale & Swain, 1980; Mitchell & Vidal, 2001). Developing language skills separately can shorten learning time and speed up the use of language skills. However, language skills must be used together for advanced applications in language teaching. For this reason, it is recommended to integrate the skills that are handled separately for the development of language skills after a certain level of competence, and to configure the language teaching accordingly at a later level (Halliday, 1978; Nunan, 1989; Widdowson, 1978).

Structuring the language teaching by grouping it according to the skills has led to a similar approach in the assessment of language skills. In order to assess the gains based on reading, listening, writing, and speaking, many tests that measure language skills are structured to consist of subtests that measure four basic skills separately. In tests designed in this way, each basic skill is accepted as a component of the language, and a score is calculated for each component as a result of the assessment (Bachman & Palmer, 1996).

In tests that evaluate four basic skills separately, test development processes specific to subtests for each skill can be followed; therefore, the approach of separating the skills according to subtests is frequently preferred. The use of skill-specific subtests has been used since the 1960s as it facilitates test development and implementation processes (Hinkel, 2010). Today, each basic language skill is measured through separate subtests within Test of English as a Foreign Language (TOEFL), Test of English for International Communication (TOEIC), International English Language Testing System (IELTS) and Pearson Test of English Academic (PTE ACADEMIC) that are used internationally to determine proficiency in various languages.

Assessment of Language Skills in Turkey

The main purpose of teaching Turkish is to make students proficient in the skill areas of their native language. It was stated in the program that language skills are related to daily life and that the development of the individual in every field is a prerequisite (MoNE, 2019a). When it is examined in detail, it is seen that the education and teaching of the Turkish language are structured on four basic language skills, which are reading, writing, listening and speaking, and grammar.

Understanding, one of the two most important aspects of the native language education and training process, is composed of listening and reading skills. Narration consists of speaking and writing skills (Kavcar, Oğuzkan & Sever, 1999). Listening and speaking skills are the skills that individuals acquire from the moment they are born and are learned before other skills. For this reason, it is aimed to support these skills in school-age children and to gain additional reading and writing skills. Unless the four basic skills are used together at a certain level, it is not possible to learn Turkish with all its functions (Doğan 2009).

Although it has an important place in the Turkish curriculum, there is no standard assessment method and assessment tool for students' four language skills in Turkish. Although there are learning outputs based on basic language skills at each grade level in the Turkish curriculum, assessment of these skills has been limited to in-class practices. In addition, no monitoring studies are conducted to assess the extent to which students have these basic skills. Language skills assessed in centralized interstage

transition examinations and periodic monitoring studies remain limited (MoNE, 2018; ÖSYM, 2018). There are subtests that assess the language skills of students in the central examinations which are applied within the scope of the High School Transition System (LGS) and Higher Education Institutions Exam (YKS), but these subtests focus only on reading skills (MoNE, 2018; ÖSYM, 2018). Turkish-Mathematics-Science Student Achievement Monitoring Study (TMF-ÖBA), which was implemented for the first time in 2019, and the Academic Skills Monitoring and Evaluation (ABİDE) focused on only the reading skill of students (MoNE, 2019b, MoNE, 2016). Additionally, the reading skill of students are assessed in international studies such as Programme for International Student Assessment (PISA) and Progress in International Reading Literacy Study (PIRLS). The results of these studies provide more important insights about students' achievements if the results are investigated in detail (Ozer, 2020).

Central examinations for assessment of basic four language skills are carried out for individuals who learn Turkish as a second language or live abroad. The Turkish Proficiency Exam (TYS) developed by Yunus Emre Institute, and the level determination and diploma exams developed by Turkish and Foreign Language Research and Application Centres (TÖMER) also assess four basic language skills. However, the target group of the examinations is individuals who learn Turkish as a foreign language. In order to assess students' basic four language skills in Turkish with standard measurement tools by overcoming this limitation, "Project for Determining and Assessing Turkish Language Proficiencies in Four Skills" was initiated by the Ministry of National Education (MoNE).

It is aimed to measure the language skills of the students within the framework of the competencies determined by the Project for Determining and Measuring Turkish Language Proficiencies in Four Skills. The results to be obtained will provide the important insights about the students competencies in language skills, language teaching and provide feedback on the effectiveness of the teaching process. Within the scope of this project, Four Skills Turkish Language Exam developed under the coordination of MoNE General Directorate of Measurement, Assessment and Examination Services. It is the first large-scale application to assess students' skills in the native language within the common assessment framework and in accordance with international assessment standards (MoNE, 2020). Language laboratories have been established in 15 provinces in order to perform the testing process at international standards. These language laboratories are equipped with headphones in which listening and recording can be performed and test cabinets that isolate external sounds.

The first step taken within the scope of the project is to develop an assessment framework to determine the scope of Turkish basic language skills. During the development of the framework, workshops were organized by the MoNE, and academics from Turkish education, experts from Turkish teaching, and measurement and evaluation specialists studied together in these workshops. Within the assessment framework developed, it was determined which behaviours to be observed in each of the basic skills, and concrete behavioural responses of language skills were developed.

The development of the assessment framework is one of the initial studies in which student behaviours to be observed within the scope of Turkish four basic language skills are determined. Although widely accepted assessment frameworks have been developed in many foreign languages, there is no framework reflecting the common view of experts in Turkish before this study. The item and task development process was carried out after the completion of the assessment framework. Each item and task developed was harmonized with the assessment framework. A pilot study of the Turkish Language Exam in Four Skills by the MoNE was conducted on 24-26 April 2019 in language laboratories with the participation of 1932 7th grade students in 15 provinces including Adıyaman, Ankara, Antalya, Aydın, Bursa, Denizli, Erzurum, Gaziantep, İstanbul, Konya, Kütahya, Muğla, Samsun, Şanlıurfa, and Trabzon. Within the scope of the test, all subtests related to four basic language skills were applied in the computerized environment.

Due to the fact that existing test applications focus only on reading skills, developed for students who learn Turkish as a second language or do not conducted as large-scale application, it is not possible to have valid and reliable data reflecting the language competencies of the students in Turkey. The number of studies focusing on determining the variables that affect the development of language skills is also

very limited for the same reason (Erkek, Batur, Kaplan & Ercan, 2017; Lüle Mert, 2013, 2014). Turkish Language Exam for Four Skills is an important step taken in order to overcome this deficiency, and the pilot study has been successfully carried out in accordance with international assessment standards. The results obtained will make it possible to implement data-based studies to develop these skills and to meet the needs of our education system by making them sustainable practices. The project outputs will provide important feedback in determining the improvements to be made in the curriculum and the development of Turkish language teaching. It will also make it possible to develop four skill tests with international standards on different levels of Turkish proficiency.

The psychometric analysis made with the data obtained from the pilot study is important in terms of ensuring that the test will be more qualifying in the initial application. Similarly, the analysis results for student characteristics on pilot study data will provide important information about the role of student characteristics in language skills. In this context, it is considered that the first results presented by the Four Skills Turkish Language Test regarding the quality of the pilot study implementation data and the relationship between student characteristics and language skills are important.

In this study, the pilot study results of the Four Skills Turkish Language Test conducted under the coordination of MoNE General Directorate of Measurement, Assessment, and Examination Services were examined, and it was aimed to determine the change of language skill performances in terms of various students, parents and school characteristics.

For this purpose, this study is conducted to answer the following research questions:

1. Is there any significant difference in students' reading, listening, writing and speaking subtest mean scores

    1 a. according to the type of school?

    1.b. according to the participation in pre-school education?

    1.c. according to the gender groups?

    1.d. according to the education levels of the parents?

2. Is there a significant relationship between students' language skills scores and 7th grade scores in Turkish, social sciences, mathematics, and science courses?

METHOD

Research Model

In the study, the current situation of the participants regarding language skills was assessed, and the relationship between language skills and various variables was examined. The descriptive correlational model was used in the design of the research. In descriptive models, phenomenon or condition in focus is examined as it is, and the current situation is described in detail (Karasar, 1999). In the descriptive correlational model, which is one of the submodels of the descriptive model, the relationships between variables are examined in detail without any external intervention.

Population and Sample

The research population is composed of the students in the seventh grade in Turkey during the academic year 2018-2019. In the sample of the study, there are 1932 seventh grade students in 15 provinces. In the sampling process, two-stage convenience sampling method was used. In this sampling type, it is possible to describe and compare the characteristics of various subgroups that are considered to be suitable according to various criteria (Büyüköztürk et al., 2016). Schools were selected according to their distance to language laboratories, type (secondary school and imam hatip secondary school), and

gender distribution criteria. After the schools were selected according to these criteria, the seventh grade classes in the school were included in the sample. In other words, all students in selected branches were applied, and after selecting the school, cluster sampling was carried out in the selection of students. In Table 1, the distribution of the study sample according to the student characteristics within the scope of the research aim is given.

Table 1. Demographic Characteristics of Students in the Study Sample

Variable	Sub Group	Frequency (f)	Ratio (%)
Gender	Female	1027	53.2
	Male	905	46.8
School Type	Secondary School	1302	67.4
	İmam Hatip Sec. School	630	32.6
Preschool Education Status	Participated	1498	77.5
	Not Participated	434	22.5
Mother's Education Level	Primary School	563	29.1
	Secondary School	287	14.9
	High School	485	25.1
	Higher Education	424	21.9
	Not Available Data	173	9.0
Father's Education Level	Primary School	304	15.7
	Secondary School	243	12.6
	High School	551	28.5
	Higher Education	663	34.3
	Not Available Data	171	8.9

As seen in Table 1, the gender distribution of the students in the study sample is quite balanced. 67.4% of the students are in secondary school, and 32.6% of them are in imam hatip secondary school. The majority of the students in the sample (77.5%) participated in pre-school education. It is determined that the ratio of students whose mothers are educated at high school or higher education level is 47%. The ratio of students whose father is educated at high school or higher education level is 62.8%.

Data Collection Tools

The data used in the study were obtained through the test battery developed for the Four Skills Turkish Language Test. Before the test battery was developed, a well-attended workshop was organized to determine the Turkish language skills to be measured, and an assessment framework was developed. Following the developing of the assessment framework, the most appropriate item and task formats were decided to assess the four language skills. A specialist group consists of Turkish linguistic experts, senior teachers in Turkish teaching practices, and measurement and evaluation experts evaluate the assessment framework and educational outputs which they expected. They agreed on item formats considering four-skill language assessment practices around the world. In this manner, it is decided to develop items related to reading and listening skills in multiple-choice format. Additionally, it is determined to develop tasks for speaking and writing skills, which enable students to structure their responses with a broader extent (MoNE, 2020). Accordingly, test-blue prints are prepared for each of the four-skills in Turkish for students in 7th grade. To decide on the cognitive levels of educational outcomes and related items, diverse taxonomies are considered, and four-level taxonomy is selected by the specialist group. In Table 2, four-level taxonomy, which is used in the pilot project, is given.

Table 2. Four-Level Taxonomy of Four Skills Turkish Language Exam

Level 1	Level 2	Level 3	Level 4
Remembering, Recognizing and Selection	Understanding and Inference (Comprehend explicitly stated information)	Inference and Interpretation (Comprehend explicitly not stated information)	Evaluation and Reflection

Educational outputs and items which are considered within the scope of the pilot project are mapped with cognitive levels in Table 2. In listening and reading sections, which consist of multiple-choice items, items are mapped with cognitive levels between level 1 and level 3 due to the limitations of item format. In the pilot project, two online test booklets for listening and reading subtests, and five online booklets for writing and speaking are developed as parallel tests. All items are developed by senior item writers in Turkish language and Turkish linguistics, and item revisions are conducted by measurement and evaluation experts. Concurrently, rubrics for open-ended tasks are developed by the specialist group, and rubrics are evaluated externally by academics from Turkish language education. Lastly, all approved items were clustered to online test booklets considering the balance of educational outcomes and item difficulties.

In the test battery, students were subjected to reading, listening, speaking, and writing subtests, respectively. The questions, tasks, and times for response according to the subtests are given in Table 3.

Table 3. Structure of Subtests in Four Skills Turkish Language Test

Subtest	Item and Task Type	Item or Task Number	Time
Reading	Multiple-Choice Item	20	30
Listening	Multiple-Choice Item	20	30
Speaking	Structured Task	2	10
Writing	Structured Task	4	60

As seen in Table 3, each of the reading and listening subtests consists of twenty multiple-choice items. In these subtests, students were given thirty minutes of response time. In the speaking subtest, students were given two tasks that were asked to explain themselves and the other to explain the steps of a process or the situation presented with the visual. Students complete this subtest in about ten minutes. In the writing subtest, students were given four tasks, including preparing a short text consisting of sentences, paragraphs, and a text including several paragraphs. The response time given to students to complete the four tasks is 60 minutes.

The subtests in the developed test battery differ structurally. Reliability analyses for reading and listening subtests consisting of multiple-choice items were performed with the Kuder-Richardson 20 coefficient frequently used in this item type. The Kuder-Richardson 20 (KR-20) coefficient is a coefficient used to calculate the internal consistency of items scored in two categories as correct and incorrect (Cronbach, 1951; Kuder & Ricardson, 1937). Kuder-Richardson 20 coefficients calculated for A and B forms of reading and listening subtests are given in Table 4.

Table 4. Internal Consistency Coefficients in Reading and Listening Subtests

Subtest	Form	Item Number	KR-20 Coefficient
Reading	A Form	20	0.720
Reading	B Form	20	0.771
Listening	A Form	20	0.768
Listening	B Form	20	0.779

As seen in Table 4, the KR-20 coefficients calculated for both forms of reading and listening subtests ranged from 0.720 to 0.779. The reliability coefficients calculated at 0.70 and above for measurement

tools used in education and psychology are considered as acceptable (Cronbach, 1951; Kuder & Richardson, 1937; Tavakol & Dennick, 2011).

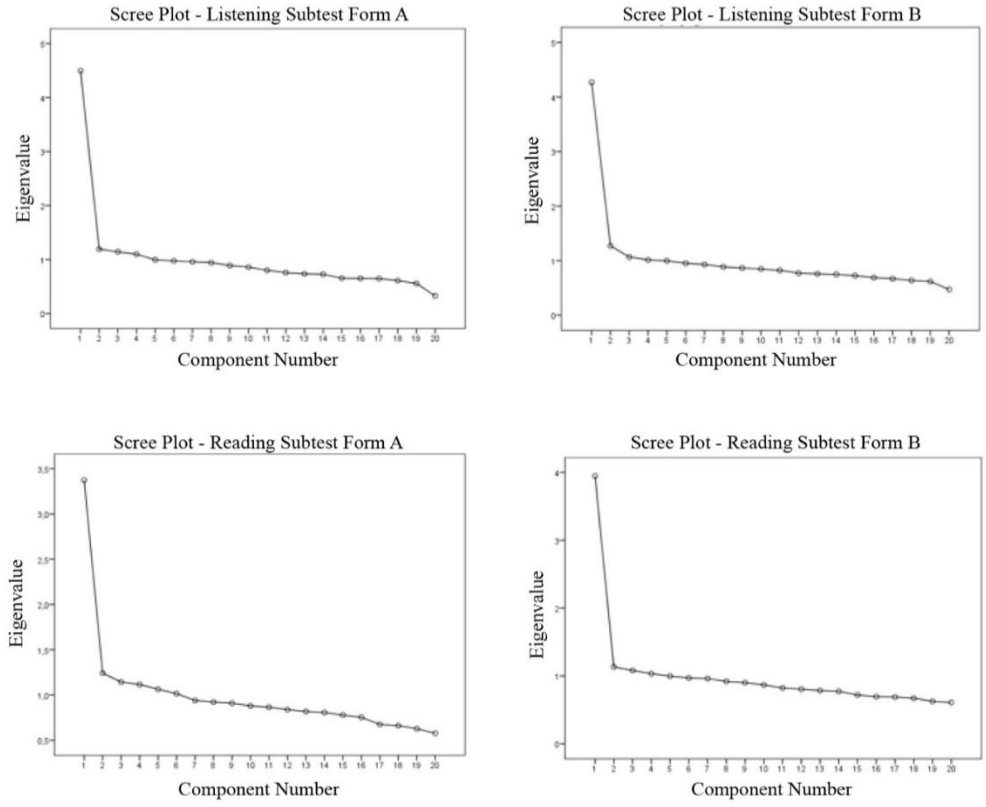
In order to provide information about the validity of the results obtained from reading and listening subtests, exploratory factor analysis (AFA) was performed to reveal the structural dimensions of both subtests. Kaiser-Meyer-Olkin (KMO) values obtained in the forms of both subtests, factor numbers with eigenvalues greater than one, variance ratio explained by the dominant factor, and factor loadings of the items under the dominant factor are given in Table 5 and scree plots are given in Chart 1.

Table 5. Exploratory Factor Analysis Results of Listening and Reading Subtests

Subtest	Form	KMO Value	Number of Factors with Eigenvalue > 1	Variance Explained by the Dominant Factor	Range of Factor Loadings
Reading	A Form	0.828	6	17%	0.047-0.532
Reading	B Form	0.888	4	20%	0.044-0.578
Listening	A Form	0.898	4	22%	0.091-0.774
Listening	B Form	0.901	4	21%	0.189-0.699

*Items below factor loading 0.30 are revised before taken into the test.

Chart 1. Scree Plots for Reading and Listening Subtests



The KMO values given in Table 5 show that the items in the forms related to the two subtests can be resolved by factor analysis. Although there are possible factors with eigenvalues greater than one in all forms, there is clearly a sharp decrease in the scree plots in Chart 1. This indicates that the items in the forms for both subtests are grouped under a single and dominant factor. The factor loadings of two of the items in each subtest form are below 0.32. The relevant items need to be strengthened in the initial

application. However, they were included in the analysis since they did not have a negative loading in this study.

In speaking and writing subtests, students are asked to answer open-ended tasks. In these subtests, students' performances are scored by assessment specialists through the answers they give in the tasks presented to them. For this purpose, assessors are trained about open-ended task assessment via rubrics, and all responses of students are assessed by assessors via a well-attended workshop conducted by MoNE. Each of the open-ended tasks is assessed by two assessors with a blinded approach, and consistency between two assessors is considered. When the score difference between assessments is significant, the final score is determined by a high-level assessor, who is a senior assessor in Turkish language education. Evidence for reliability in these types of tests is mostly provided by the interrater reliability method. In this method, the consistency between the scores given by the raters for the answers of students to the tasks is examined (Crocker & Algina, 1986). Cramer's V coefficient (Cramer's V) and contingency coefficient were used to obtain evidence of inter-rater reliability, and the coefficients with regard to speaking subtest were given in Table 6 based on forms.

Table 6. Consistency Coefficients between Raters in the Speaking Subtest *

Form	Task Type	Cramer's V (Mean)	Contingency Coefficient (Mean)
A Form	First Task	0.56	0.68
	Second Task	0.58	0.69
B Form	First Task	0.47	0.62
	Second Task	0.66	0.75
C Form	First Task	0.56	0.69
	Second Task	0.56	0.67
D Form	First Task	0.59	0.69
	Second Task	0.62	0.69
E Form	First Task	0.49	0.58
	Second Task	0.46	0.56

* *In the first task, students are asked to introduce themselves, in the second task to explain the steps of a process or the situation presented with the visual.*

As seen in Table 6, in the speaking subtest, the V coefficients calculated between the raters were between 0.46 and 0.66, and the contingency coefficients were between 0.56 and 0.75. There are no generally accepted standards as in the other types of reliability for the V coefficient and the contingency coefficient, whose values vary between 0 and 1. However, V coefficients greater than 0.25 are considered to provide information about the general agreement between the two variables (Akoğlu, 2018). It is seen that the V coefficients given in Table 4 are well above this criterion. The contingency coefficients calculated in this subtest are higher than the V coefficients and indicate that the consistency between raters is relatively high.

The reliability coefficients between the raters calculated in the writing subtest are given in Table 7.

Table 7. Interrater Consistency Coefficients Calculated in Writing Subtest *

Form	Task Type	Cramer's V (Mean)	Contingency Coefficient (Mean)
A Form	First Task	0.93	0.85
	Second Task	0.81	0.81
	Third Task	0.60	0.72
	Fourth Task	0.68	0.74
B Form	First Task	0.86	0.83
	Second Task	0.71	0.71
	Third Task	0.56	0.66
	Fourth Task	0.70	0.76
C Form	First Task	0.91	0.85
	Second Task	0.80	0.81
	Third Task	0.66	0.75
	Fourth Task	0.69	0.77
D Form	First Task	1.00	0.87
	Second Task	1.00	0.82
	Third Task	0.68	0.75
	Fourth Task	0.66	0.74
E Form	First Task	0.83	0.82
	Second Task	0.92	0.68
	Third Task	0.59	0.69
	Fourth Task	0.74	0.71

**Students are expected to write a sentence in the first and second tasks, a paragraph in the third task, and a text composed of several paragraphs in the fourth task.*

As seen in Table 7, the mean V coefficients calculated in different forms of the writing subtest ranged from 0.56 to 1, and the mean consistency coefficients ranged from 0.66 to 0.85. These coefficients indicate a high level of consistency among raters, as in the speaking subtest.

Evidence regarding the reliability and validity of the results obtained from the test battery shows that the data obtained from the pilot study is sufficient in terms of psychometric perspective. As it can be seen from Table 4, Table 5, and Table 6, it is possible to revise particular items and tasks in the test battery to be more qualified in the initial application, but in this study, all items and tasks are included in the analysis in their current form.

Ethics Committee Permission

The data of this research were used with the letter number of 42497731-605.99-E.6452557 dated 17.04.2020 of the General Directorate of MEB Measurement, Evaluation and Examination Services.

Data Analysis

In the study, t-test, single-factor variance analysis (ANOVA), effect size analysis, Pearson correlation analysis were used for the analysis of quantitative data obtained with the test battery. The t-test and single-factor ANOVA were used to examine the significance of the difference between the language skill mean scores of the groups, and the eta-square effect size was used to analyze the effect of the variables on the scores. The Pearson correlation analysis was used to determine the and significancy and strength of the relationship between the variables. Significant differences between the groups were interpreted by taking into account their effect size. Criteria for effect size (partial eta-square) are as follows: PES<0.02 is small, 0.02<PES<0.13 is medium, and PES<0.13 is a high level of effect (Miles & Shelvin, 2001).

Participation in pre-school education, gender, and parents' level of education are selected as possible effective variables on the language skill of students. It is shown that these demographic and educational

variables lead to significant changes on students' language development and skills (Bakken, Brown & Downing, 2015; Catts, Fey, Zhang & Tomblin, 2001; Reilly, Neuman & Andrews, 2019; Schermse et al., 2018, Storch & Whitehurst, 2002). The difference between school types is also examined to have insights about the possible effect of educational program differences on language skills of students.

RESULTS

In the findings section of the research, descriptive statistics, and the findings related to each research question are given, respectively.

The mean scores and other descriptive statistics obtained by students in subtests for language skills are given in Table 8.

Table 8. Descriptive Statistics of Language Skill Subtest Scores

Subtest	Possible Score Range	Lowest Score	Highest Score	$\bar{X}$	SD
Reading	0-20	0	20	10.63	3.63
Listening	0-20	0	20	11.70	2.98
Writing	0-36	1	36	16.82	8.09
Speaking	0-36	15	36	27.21	3.95

As seen in Table 8, the mean scores calculated in the reading and listening subtests, where the scores that can vary between 0 and 20, are quite close. In the writing and speaking subtests ranging from 0 to 36, the students perform quite differently. It can be seen in Table 7 that the students perform relatively high in the speaking subtest (X = 27.21, SS = 3.95) and that they showed relatively low performance in the writing subtest (X = 16.82, SS = 8.09).

Findings Related to the First Research Question

The findings of the t-test and effect size related to the research question 'is there any significant difference in students' reading, listening, writing and speaking subtest mean scores according to the type of school?' are shown in Table 9.

Table 9. t-Test Results of Language Skill Subtest Scores by School Type

Subtest	School Type	n	$\bar{X}$	SD	df	*t*	η^2
Reading	Secondary School	1238	10.56	3.63	1832	1.058	---
	İmam Hatip Sec. School	596	10.78	3.62			
Listening	Secondary School	1240	11.73	2.99	1836	0.555	---
	İmam Hatip Sec. School	598	11.65	2.99			
Writing	Secondary School	1024	16.64	8.05	1523	0.883	---
	İmam Hatip Sec. School	501	17.18	8.26			
Speaking	Secondary School	677	27.24	3.98	1019	0.276	---
	İmam Hatip Sec. School	344	27.15	3.83			

As can be seen from the t-test results in Table 9, the type of school has not a significant effect on the language skills of the students. In other words, the students who attend secondary school and imam hatip secondary school have a similar level of scores in reading, listening, writing, and speaking subtests. The effect sizes show that the effect of school type on students' language skills is negligible.

As seen from the t-test results, there is no significant difference between the reading subtest mean scores by school type ($t_{(1832)} = 1.058$, $p>0.05$). In the reading subtest of the students in imam hatip secondary

school, the mean score is calculated as 10.78, and the mean score of the students in secondary schools is 10.56. It is observed that the effect of school type on reading subtest scores is negligible.

It is observed that mean listening subtest scores given by school type are quite close to each other, and students in diverse secondary school types perform similarly in this subtest. There is no significant difference between the listening subtest scores of the students according to school type ($t_{(1836)} = 0.555$, $p>0.05$). The effect size analysis also showed that the school type does not have a significant effect on the listening subtest scores.

In the listening subtest, the mean score of the students who are in imam hatip secondary school is 17.18, and that of the students who are in secondary schools is 16.64. As can be seen from the t-test results, there is no significant difference between the mean scores of the students in both school types in writing subtest ($t_{(1523)} = 0.883$, $p>0.05$). As a result of the effect size analysis, it is shown that the school type does not have a significant effect on writing subtest scores.

The mean speaking subtest score of the students who are in imam hatip secondary school is calculated as 27.15. The mean subtest score of the students in secondary school is 27.24. According to the t-test test results, there is no significant difference between the mean scores of the speaking subtest by school type ($t_{(1019)} = 0.276$, $p>0.05$). The result of the effect size analysis shows that the school type does not have a significant effect on the speaking subtest scores.

The findings of the t-test to find the answer to the research question "Is there any significant difference in students' reading, listening, writing and speaking subtest mean scores according to the participation to pre-school education' are shown in Table 10 together with descriptive statistics.

Table 10. t-Test Results of Language Skill Subtest Scores According to Preschool Education Status

Subtest	Pre-School Education Status	n	$\underline{X}$	SD	df	*t*	η^2
Reading	Par. Pre-School Edu.	1419	10.87	3.58	1832	6.328*	.021
	Not Par. Pre-School Edu.	415	9.60	3.62			
Listening	Par. Pre-School Edu.	1422	11.91	2.92	1836	5.624*	.017
	Not Par. Pre-School Edu.	416	10.99	3.10			
Writing	Par. Pre-School Edu.	1171	17.38	8.24	1523	5.272*	.018
	Not Par. Pre-School Edu.	354	14.81	7.37			
Speaking	Par. Pre-School Edu.	832	27.22	3.94	1019	0.313	---
	Not Par. Pre-School Edu.	189	27.32	3.88			

*$p<0.05$.

The t-test results given in Table 10 show that participation in pre-school education leads to a significant difference in all subtest scores except speaking. Therefore, the reading, listening, and writing subtest scores of students who participate in preschool education are significantly higher. It is seen that participating preschool has its strongest effect on reading skill. As can be seen from the t-test results, there is a significant difference between the mean reading skills scores of students according to their preschool education status ($t_{(1832)} = 6.328$, $p<0.05$, $n^2= 0.021$). Students who participate in preschool education have a higher mean score in the reading subtest. Effect size analysis shows that preschool education has a significant effect on reading subtest scores, but this size of effect is small.

The mean listening score of students who did not participate in pre-school education is calculated as 10.99 in this subtest. The mean listening score of students receiving preschool education is 11.91. There is a significant difference between the mean listening subtest scores of the students according to their pre-school education status ($t_{(1836)} = 5.624$, $p<0.05$, $n^2= 0.017$). The mean listening subtest score of the students who participated in preschool education is higher. According to the results of the effect size, it was determined that the effect of participating in preschool education on listening subtest scores was low.

The mean writing subtest score of students who did not participate in preschool education is 14.81. The mean score of the students who participated in preschool education in the writing subtest is calculated as 17.38. The t-test results show that students who participated in preschool education have significantly higher scores in writing subtest than students who did not participate in preschool education ($t_{(1523)}$ = 5.272, p<0.05, n^2= 0.018). The effect size results show that preschool education has a low impact on students' writing subtest scores.

The mean score in the speaking subtest of students who did not participate in preschool education is calculated as 27.32. The mean score of the students who participated in preschool education is 27.22. The t-test results show that preschool education does not lead to a significant difference between the mean speaking scores ($t_{(1019)}$ = 0.313, p>0.05).

The findings of the t-test to answer the research question of 'is there any significant difference in students' reading, listening, writing and speaking subtest mean scores according to the gender groups?' are presented in Table 11 together with descriptive statistics.

Table 11. t-Test Results of Language Skill Subtest Scores by Gender

Sub Test	Gender	n	$\underline{X}$	SD	df	*t*	η^2
Reading	Female	971	10.92	3.57	1832	4.163*	.009
	Male	863	10.21	3.66			
Listening	Female	975	12.00	2.79	1836	4.512*	.011
	Male	863	11.37	3.17			
Writing	Female	812	18.32	8.12	1523	8.055*	.041
	Male	713	15.03	7.76			
Speaking	Female	561	27.96	3.79	1019	6.618*	.041
	Male	460	26.36	3.92			

**p<0.05*.

According to the results given in Table 11, the effect of gender on language skills leads to a significant difference in all subtests. It was determined that the mean scores of female students in all reading, listening, writing, and speaking subtests are significantly higher than male students. The effect size analysis shows that the difference between the mean scores of female and male students is even greater in writing and speaking subtests. According to the t-test results related to the reading subtest scores, there is a significant difference between the mean scores of male and female students ($t_{(1832)}$ = 4.163, p<0.05, n^2= 0.009). Female students' mean reading scores are higher than male students. According to the results of the effect size analysis, the effect of gender on the reading subtest scores is low.

In the listening subtest, the mean score of male students is calculated as 11.37, and the mean score of female students is 12. According to the t-test results related to the listening subtest scores, there is a significant difference between the mean scores of female and male students ($t_{(1836)}$ = 4.512, p<0.05, n^2= 0.011). Listening mean scores of female students are significantly higher than male students. In the effect size analysis, it is observed that the effect of gender on reading scores is low.

As can be seen from the t-test results, there is a significant difference between female students' mean writing score and male students' mean writing score ($t_{(1523)}$ = 8.055, p<0.05, n^2= 0.041). Female students' mean writing scores are higher than male students. It is determined that the effect of gender on writing scores is low.

The mean score of male students in the speaking subtest is calculated as 26.36 and female students as 27.96. The mean score of female students in the speaking subtest is significantly higher than the mean of the male students ($t_{(1019)}$ = 6.618, p<0.05, n^2= 0.041), but the effect of gender on the speaking subtest is found to be low.

The single-factor ANOVA findings to the research question "Is there any significant difference in students' reading, listening, writing, and speaking subtest mean scores according to the education levels of the mothers?' are shown in Table 12 together with descriptive statistics.

Table 12. ANOVA Results of Language Skill Scores According to Mother's Education Level

Subtest	Education Level	n	$\underline{X}$	SD	df	*F*	η^2
Reading	Primary Sch.	543	9.27	3.56	3	67.817*	.109
	Secondary Sch.	277	10.00	3.55			
	High School	461	11.88	3.21			
	Higher Edu.	386	12.86	3.41			
Listening	Primary Sch.	545	10.84	3.04	3	38.569*	.065
	Secondary Sch.	277	11.43	2.89			
	High School	461	11.88	2.70			
	Higher Edu.	388	12.87	2.91			
Writing	Primary Sch.	451	14.88	7.80	3	20.131*	.041
	Secondary Sch.	232	16.77	8.24			
	High School	382	16.47	7.58			
	Higher Edu.	324	19.37	8.36			
Speaking	Primary Sch.	275	26.03	4.14	3	17.957*	.056
	Secondary Sch.	144	26.92	3.80			
	High School	264	27.59	3.72			
	Higher Edu.	237	28.45	3.67			

*$p<0.05$.

As can be seen from the ANOVA results in Table 12, the education level of the mother leads to a significant difference in all subtest scores. In other words, students whose mothers graduated from higher education have significantly higher reading, listening, writing, and speaking scores. It is observed that the education level of the mother has its greatest impact on reading scores.

There is a significant difference between the mean reading scores of the students according to the education level of the mother ($F_{(3,1667)} = 67.817$, $p<0.05$, $n^2= 0.109$). As the education level of the mother increases, students' mean scores in the reading subtest increase. According to the results of the effect size analysis, the mother education level has a small effect on reading scores.

It is seen that the education level of the mothers leads to a significant difference in the listening scores ($F_{(3,1671)} = 38.569$, $p<0.05$, $n^2= 0.065$). It is determined that the mean score of the students whose mothers are graduates of higher education is 12.87 in the listening subtest, and the mean of the students whose mothers are primary school graduates is 10.84. As a result of the effect size analysis, it is determined that the effect of mother education level is low on the listening scores.

The mean of the writing subtest scores of the students whose mothers are higher education graduates is 19.37, and those whose mothers are graduated from primary school are calculated as 14.88. ANOVA results show that students whose mothers have higher education levels have significantly higher scores than other students' scores ($F_{(3,1389)} = 20.131$, $p<0.05$, $n^2= 0.041$). The effect of mother education level on students' writing subtest scores is examined, and it is showed that this effect is low.

In line with the ANOVA results, it was observed that the level of mother education leads to a significant difference between the students' speaking scores ($F_{(3,920)} = 17.957$, $p<0.05$, $n^2= 0.056$). As a result of the effect size analysis, it is determined that the effect of mother education level on students' speaking scores is low.

The findings of the single-factor ANOVA to find the answer to the research question 'is there any significant difference in students' reading, listening, writing, and speaking subtest mean scores according to the education levels of the fathers? are shown in Table 13, together with descriptive statistics.

Table 13. ANOVA Results of Language Skill Subtest Scores According to Father's Education Level

Subtest	Education Level	n	$\underline{X}$	SD	df	*F*	η^2
Reading	Primary Sch.	296	8.91	3.60	3	61.218*	.100
	Secondary Sch.	233	9.70	3.47			
	High School	531	10.38	3.44			
	Higher Edu.	611	11.94	3.35			
Listening	Primary Sch.	296	10.48	3.10	3	40.270*	.067
	Secondary Sch.	234	11.33	2.78			
	High School	532	11.54	2.90			
	Higher Edu.	613	12.61	2.98			
Writing	Primary Sch.	249	14.27	7.84	3	20.458*	.042
	Secondary Sch.	194	15.46	7.44			
	High School	431	16.48	7.87			
	Higher Edu.	514	18.69	8.20			
Speaking	Primary Sch.	139	26.25	4.12	3	11.019*	.035
	Secondary Sch.	126	26.43	3.86			
	High School	300	27.01	3.93			
	Higher Edu.	358	28.09	3.70			

**p<0.05*.

As seen in Table 13, the father's education level leads to a significant difference in all subtest scores. Therefore, students whose father graduated from a higher education level have higher reading, listening, writing, and speaking subtest scores. It is shown that the education level of the father has its strongest effect on reading scores.

According to ANOVA results, it is seen that the father's education level leads to a significant difference in mean reading scores of the students ($F_{(3.1671)} = 61.218$, $p<0.05$, $n^2= 0.100$). Effect size results showed that the father's education level has a small effect on students' reading scores.

It is shown that the father's education level also leads to a significant difference in the mean listening scores of students ($F_{(3.1675)} = 40.270$, $p<0.05$, $n^2= 0.067$). Effect size results showed that the father's education level has a small impact on students' listening scores.

According to the ANOVA results, there is a significant difference between the mean writing scores according to the father's education level ($F_{(3.1388)} = 20.458$, $p<0.05$, $n^2= 0.042$). The mean of the speaking scores of students whose father graduated from higher education is calculated as 18.69, and the mean of the speaking scores of the students whose father graduated from primary school is calculated as 14.27. The results of the effect size analysis showed that the level of father's education has a low impact on students' writing scores.

The mean speaking score of the students whose fathers are higher education graduates is calculated as 28.09, and the mean score of the students whose fathers are primary school graduates is calculated as 26.25. The results show that there is a significant difference between the students' mean speaking score according to the education level of the father ($F_{(3.923)} = 11.019$, $p<0.05$, $n^2= 0.035$). The results of the effect size analysis showed that the level of father's education has a low impact on students' speaking scores.

Findings for the Second Research Question

The relationship between the students' reading, listening, writing and speaking scores and their scores in Turkish, social sciences, mathematics and science courses was analyzed via Pearson correlation coefficient to answer the research question "*Is there a significant relationship between the students' language skills scores and the 7th grade scores in Turkish, social sciences, mathematics, and science?*" and the findings are shown in Table 14.

Table 14. The Relationship between Students' Reading, Listening, Writing, Speaking Scores and Their Scores in Turkish, Social Sciences, Mathematics and Science Courses in 7th Grade*

Subtest	Course Score	r
Reading	Turkish	0.66*
	Social Studies	0.64*
	Mathematics	0.61*
	Science	0.62*
Listening	Turkish	0.53*
	Social Studies	0.53*
	Mathematics	0.49*
	Science	0.50*
Writing	Turkish	0.38*
	Social Studies	0.35*
	Mathematics	0.35*
	Science	0.33*
Speaking	Turkish	0.29*
	Social Studies	0.26*
	Mathematics	0.28*
	Science	0.24*

* *p<0.05*

As can be seen in Table 14, there are significant relationships between all four subtests of Four Skills Turkish Language Test and scores of Turkish, social sciences, mathematics, and science courses in 7th grade. Correlation coefficients calculated at the subtest level are explained below.

According to the results of Pearson correlation analysis given in Table 14, there is positive, statistically significant relationships between the students' reading subtest scores and scores of Turkish ($r = 0.66$, $p < 0.05$), social sciences ($r = 0.64$, $p < 0.05$), mathematics ($r = 0.61$, $p < 0.05$) and science courses ($r = 0.62$, $p < 0.05$). These results show that the performances of the students in the reading subtest and their performances in all four courses are significantly related.

There is positive, statistically meaningful and medium level relationships between the students' listening scores and scores of Turkish ($r = 0.53$, $p < 0.05$), social sciences ($r = 0.53$, $p < 0.05$), mathematics ($r = 0.49$, $p < 0.05$) and science courses ($r = 0.50$, $p < 0.05$). These results show that the performances of the students in the listening subtest and their performances in all four courses are significantly related.

There is positive, significant and medium-level relationships between the students' writing scores and scores of Turkish ($r = 0.38$, $p < 0.05$), social sciences ($r = 0.35$, $p < 0.05$), mathematics ($r = 0.35$, $p < 0.05$) and science courses ($r = 0.33$, $p < 0.05$). According to these findings, the students' writing scores are significantly correlated with their performance in all four courses. The correlation between the subtest scores and the course scores is found to be higher in the writing subtest than in speaking subtest.

There is positive, significant and low-level relationships between the students' speaking scores and scores of Turkish ($r = 0.29$, $p < 0.05$), social sciences ($r = 0.26$, $p < 0.05$), mathematics ($r = 0.28$, $p < 0.05$) and science courses ($r = 0.24$, $p < 0.05$). These results show that the scores of the students in the speaking are significantly related to their performance in all four courses, but the level of relationship between them is low.

CONCLUSION AND DISCUSSION

Language skill is one of the basic skills that individuals must have in order to express themselves and be a part of the society. It has been shown in academic studies that the individual's competencies in the native language and many educational outcomes are related, especially academic achievement (Akbaşlı, Şahin & Yaykıran, 2016; Mahmud, 2014; Shali, 2017). Therefore, language skills have an important

role in the social and academic life of individuals. Because of this role, language skills are among the most important skills acquired through education.

The ways and methods in the development of language skills have also influenced the methods used to evaluate these skills. Approaches in which the four skills are assessed and scored separately in the assessment of native language and foreign language skills are the majority. Today, these basic skills are assessed separately in exams such as TEOFL, TOEIC, IELTS, PTE, which are often used for qualification.

In the Turkish language teaching program as the native language, there are educational outcomes to improve students' basic four language skills. However, there is no standard measurement method to assess the extent to which students have these basic skills, and no monitoring study is available on this subject. In interstage transition examinations such as LGS and YKS, and periodic monitoring studies as TMF-ÖBA and ABİDE focus only on reading skills. In this context, detailed data on students' listening, writing, and speaking skills are not available. In order to overcome this important deficiency, MoNE developed the Four Skills Turkish Language Test in 2019, and the pilot study is conducted under the coordination of the General Directorate of Measurement, Evaluation, and Examination Services.

The results indicate that students performed relatively high in the speaking subtest and relatively low in the writing subtest. It is observed that female students had higher mean scores than male students in all subtests. It can be stated that this result is consistent with the results of inter-stage examinations and monitoring studies in particular for the reading subtest (MoNE, 2018, MoNE2019b, ÖSYM, 2018). Findings are in coherence with the results of the monitoring study examined by Reilly, Neuman, and Andrews (2019). There are additional findings on internationally applied TOEFL and NAEP exams that female students are more successful, but the difference between gender groups is small (ETS, 2001; ETS, 2017). The fact that female students are performing better than male students in some language skills is seen in Dutch (van der Silk, van Hout & Schepens, 2015). Findings are also consistent with the PISA 2018 application that female students are performing better than male students in the reading field in the sample of Turkey (MoNE, 2019c). This finding shows that students from different gender groups may have diverse levels of linguistic skills.

It is determined that the mean scores of students attending secondary school and imam hatip secondary school did not show any significant difference in any subtest. In other words, the type of school the student attends does not have a significant effect on students' language skills. These findings are in coherence with that the graduates enrolled in imam hatip secondary school and other secondary schools according to the 2018 LGS central exam results (MoNE, 2018). Similarly, within the scope of 8th grade application of ABİDE 2016, it is determined that mean scores of imam hatip secondary school and secondary school students are quite close (MoNE, 2016). According to the results, students in two school types performed at the same level in their listening, writing, reading, and speaking skills.

It is determined that the students who participated in preschool education show higher performance than the students who did not participate in preschool education in all subtests except speaking. Considering the effect of preschool education on language development, these findings are seen to be consistent with the literature (Bakken, Brown & Downing, 2015; Schermse et al., 2018). It has been demonstrated in academic research that providing students with verbal skills through education in the preschool period positively contributes to language development and the psychological development of the individuals (Catts, Fey, Zhang, & Tomblin, 2001; Storch & Whitehurst, 2002). In speaking skills, why preschool education does not have a significant effect requires a more detailed investigation as a separate research subject.

Another finding is that the increase in parents' education level also increases the mean scores of the students in all subtests. The fact that parents from higher education levels use comparatively higher level of intellectual and complex language in home and read more with their children (Raikes et al, 2006; Rowe, Pan & Ayoub, 2005; Tamis-Lemonda & Rodriguez, 2009) is a possible reason for this significant difference between students. As parents' level of education is one of the components of students' social

background, and social background has a significant impact on students' academic achievement (Ozer & Perc, 2020; Schuetz, Ursprung & Woessmann, 2008), it is expected that students' language skills are positively correlated with parents' level of education. It is determined that the effect of mother and father education level on students' listening skills is higher than their gender and participation in preschool education. The findings are consistent with inter-stage examination results (MoNE, 2018) and academic studies abroad (Khodadady & Alaee, 2012; Richels, Johnson, Walden & Conture, 2013).

The study also examined the relationship between language skills and students' Turkish language, social sciences, mathematics, and science course scores. It has been determined that reading, listening, writing, and speaking skills have a significant relationship between the scores of four courses in levels ranging from low to medium (r=0.24 - r=0.66). This finding, which is important information about the validity of the study, also revealed an important implementation regarding the assessment of language skills in classrooms. The relationship between the reading scores and scores of the four courses is quite higher than the other language skills (between r = 0.61 - r = 0.66). One possible reason for this is that reading skills are used intensively in classroom assessments.

Findings obtained within the scope of the pilot study show that the test battery will make important contributions to the assessment of students' basic language skills. The findings support the validity of the pilot study, provide sufficient psychometric evidence, and the findings are supported by language development and assessment literature. The results of the future initial study will provide important feedback for native language teaching. Findings of the pilot study of the 'Four Skills Turkish Language Test' conducted by MoNE for the first time show that the assessment framework and data provide valid and reliable findings as a whole. Based on the data from the pilot study, it will be possible to develop certain levels of exams at the same standards with international qualifications in four skills of Turkish, both to strengthen native language education in schools and to make a more detailed analysis and to enhance educational processes.

REFERENCES

Akbaşlı, S., Şahin, M., & Yaykıran, Z. (2016). The effect of reading comprehension on the performance in science and mathematics. *Journal of Education and Practice, 7*(6), 108-121.

Akoğlu, H. (2018). User's guide to correlation coefficients. *Turk J Emerg Med,18*(3), 91–93.

Akram, A., & Malik, A. (2010). Integration of language learning skills in second language acquisition. *International Journal of Arts and Sciences, 3*(14), 231 – 240.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests.* USA: Oxford University Press.

Bakken, L., Brown, N., & Downing, B. (2015). Early childhood education: The long-term benefits. *Journal of Research in Childhood Education, 31*(2), 255-269.

Barın, M. (2000). Yabancı dil öğretiminde konuşma becerisinin önemi. *Atatürk Üniversitesi Sosyal Bilimler Dergisi*, 26, 123-127.

Bayyurt, Y. (2013). Current perspectives on sociolinguistics and English language education. *The Journal of Language Teaching and Learning, 1*, 69-78.

Boonkit, K. (2010). Enhancing the development of speaking skills for non-native speakers of English. *Procedia Social and Behavioral Sciences, 2*, 1305–1309.

Brown, H. D. (2001). *Teaching by principles: An interactive approach to language pedagogy* (2nd ed.). New York: Pearson Education.

Budria, S., Colino, A., & Martined de Ibarreta, C. (2019). The impact of host language proficiency on employment outcomes among immigrants in Spain. *Empirica, 46*, 625–652.

Büyüköztürk, Ş., Çakmak, E. K., Akgün, Ö. E., Karadeniz, Ş., & Demirel, F. (2016). *Bilimsel araştırma yöntemleri.* Ankara: Pegem Akademi Yayınları.

Canale, M., & Merrill, S. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*(1), 1-47.

Catts, H. W., Fey, M.E., Zhang, X., & Tomblin, B. (2001). Estimating the risk of future reading difficulties in kindergarten children: A research-based model and its clinical implementation. *Language, Speech, and Hearing Services in Schools, 32*, 38–50.

Chengyu, N. (2018). Implications of interrelationship among four language skills for high school English teaching. *Journal of Language Teaching and Research, 9*(2), 418-423.

Cook, V. J. (1999). Going beyond the native speaker in language teaching. *TESOL Quarterly, 33*(2), 185-209.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* New York: Holt, Rinehart and Winston.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297-334.

Deniz, K. (2000). *Yazılı anlatım becerileri bakımından köy ve kent beşinci sınıf öğrencilerinin durumu*. (Yayınlanmamış yüksek lisans tezi). Çanakkale On Sekiz Mart Üniversitesi Sosyal Bilimler Enstitüsü, Çanakkale.

Doğan, Y. (2009). Konuşma becerisinin geliştirilmesine yönelik etkinlik önerileri. *Türk Eğitim Bilimleri Dergisi*, 7(1), 185-204.

Ergin, M. (1998). *Türk dil bilgisi.* İstanbul: Bayrak Yayınları.

Erkek, G., Batur, Z., Kaplan, K., & Ercan, E. (2017). Türkçe eğitimi ve yabancı dil öğretiminde dört temel dil becerisinin edinimine ilişkin öğretmen görüşleri. *Avrasya Dil Eğitimi ve Araştırmaları Dergisi, 1*(1), 42-75.

Educational Testing Service (2001). *Differences in the gender gap: Comparisons across racial/ethnic groups in education and work*. Policy Information Report. Princeton: Education Testing Service.

Educational Testing Service (2017). *Test and score data: Score and data summary for TOEFL iBT Test*. Princeton: Education Testing Service.

Gautam, P. (2019). Integrated and segregated teaching of language skills: An exploration. *Journal of NELTA Gandaki (JoNG), 1,* 100-107.

Gazzola, M., & Mazzacani, D. (2019). Foreign language skills and employment status of European natives: Evidence from Germany, Italy and Spain. *Empirica, 46,* 713–740.

Halliday, M. A. K. (1978). *Language as social semiotic: The social interpretation of language and meaning.* London: Edward Arnold.

Hinkel, E. (2002). *Second language writers' text: Linguistic and rhetorical.* Mahwah: Lawrence Erlbaum.

Hinkel, E. (2010). Integrating the four skills: Current and historical perspectives. In Robert B. Kaplan. (Ed.), *The Oxford handbook of applied linguistics* (2nd ed.) (pp. 110-125).

Jing, W.U. (2006). Integrating skills for teaching EFL: Activity design for the communicative classroom. *Sino-US English Teaching, 3*(12).

Kantemir, E. (1997). *Yazılı ve sözlü anlatım*. Ankara: Engin Yayınevi.

Karasar, N. (1999). *Bilimsel araştırma yöntemi.* Ankara: Nobel Yayınevi.

Kavcar C., Oğuzkan F., & Sever, S. (1995). *Türkçe öğretimi (Türkçe ve sınıf öğretmenleri için)*. Ankara: Engin Yayınevi.

Khodadady, E., & Alaee, F. F. (2012). Parent education and high school achievement in English as a foreign language. *Theory and Practice in Language Studies, 2*(9), 1811-1817.

Kuder, G. F., & Richardson, M. W. (1937). The theory of estimation of test reliability. *Psychometrika. 2,* 151-160.

Lüle Mert, E. (2013). Türkçe öğretmen adaylarının dört temel dil becerisine ilişkin algılarının metaforlar aracılığıyla analizi. *Uluslararası Sosyal Araştırmalar Dergisi, 6*(27), 357-372.

Lüle Mert, E. (2014). Türkçenin eğitimi ve öğretiminde dört temel dil becerisinin geliştirilmesi sürecinde kullanılabilecek etkinlik örnekleri. *Ana Dili Eğitimi Dergisi, 2*(1), 23-48.

Mahmud, M. M. (2014). Communication aptitude and academic success. *Procedia - Social and Behavioral Sciences, 134,* 125 – 133.

Manaj-Sadiku, L. (2015). The importance of four skills reading, speaking, writing, listening in a lesson hour. *European Journal of Language and Literature Studies, 1*(1), 29-31.

Ministry of National Education (2016). *Akademik becerilerin izlenmesi ve değerlendirilmesi: 8. sınıf raporu.* Ankara: MEB Yayınları.

Ministry of National Education (2018). *2018 Liselere geçiş sistemi: Merkezi sınavla yerleşen öğrencilerin performansı.* Eğitim Analiz ve Değerlendirme Raporları Serisi No:3. Ankara: MEB Yayınları.

Ministry of National Education (2019a). *Türkçe dersi öğretim programı*. Ankara: MEB Yayınları.

Ministry of National Education (2019b). *Türkçe-Matematik-Fen Bilimleri öğrenci başarı izleme araştırması (TMF-ÖBA)-I: 2019 4. sınıf seviyesi.* Eğitim Analiz ve Değerlendirme Raporları Serisi No:9. Ankara: MEB Yayınları.

Ministry of National Education (2019c). *PISA 2018 Türkiye ön raporu.* Eğitim Analiz ve Değerlendirme Raporları Serisi No:10. Ankara: MEB Yayınları.

Ministry of National Education (2020). *Dört beceride Türkçe dil sınavı: Pilot çalışma sonuçları.* Eğitim Analiz ve Değerlendirme Raporları Serisi No:11. Ankara: MEB Yayınları.

Miles, J., & Shevlin, M. (2001). *Applying regression and correlation: A guide for students and researchers.* London: Sage Publishing.

Mitchell, C. B., & Vidal, K. E. (2001). Weighing the ways of the flow: Twentieth century language instruction. *Modern Language Journal, 85*, 26-38.

Nunan, D. (1989). Designing *tasks for the communicative classroom.* Cambridge: Cambridge University Press.

Ölçme, Seçme ve Yerleştirme Merkezi (2018). *2018 YKS değerlendirme raporu*. Değerlendirme Raporları Serisi No: 9. Ankara: ÖSYM Yayınları.

Ozer, M., & Perc, M. (2020). Dreams and realities of school tracking and vocational education. *Palgrave Communications, 6,* 34.

Ozer, M. (2020). What PISA tells us about performance of education systems?. *Bartın University Journal of Faculty of Education, 9*(2), 217-228.

Özkırımlı, A. (1994). *Dil ve anlatım*. Ankara: Ümit Yayınevi.

Perc, M., Ozer, M., & Hojnik, J. (2019). Social and juristic challenges of artificial intelligence. *Palgrave Communications, 5,* 61.

Raikes, H., Pan, B. A., Luze, G., Tamis-LeMonda, C.S., Brooks-Gunn, J., Constantine, J., Tarullo, L. B., Raikes, H.A., & Rodriguez, E.T. (2006). Mother-child bookreading in low-income families: Correlates and outcomes during the first three years of life. *Child Development, 77*(4), 924-953.

Reilly, D., Neumann D. L., & Andrews, G. (2019). Gender differences in reading and writing achievement: Evidence from the National Assessment of Educational Progress (NAEP). *American Psychologist, 74*(4), 445– 458.

Richels, C. G., Johnson, K. N., Walden, T. A., & Conture, E. G. (2013). Socioeconomic status, parental education, vocabulary and language skills of children who stutter. *J Commun Disord., 46*(4), 361-74.

Rowe, M. L., Pan, B. A., & Ayoub, C. (2005). Predictors of variation in maternal talk to children: A longitudinal study of low-income families. *Parenting: Science and Practice, 5*(3), 259-283.

Schmerse, D., Yvonne A., Flöter, M., Wieduwilt, N., Rossbach, H. G., & Tietze, W. (2018). Differential effects of home and preschool learning environments on early language development. *British Educational Research Journal, 44*(2), 338-357.

Schuetz, G., Ursprung, H. W., & Woessmann, L. (2008). Education policy and equality of opportunity. *Kyklos 61*(2), 279–308.

Shali, S. K. (2017). The power of listening ability and its effects on academic performance: An examination of college students. *Imperial Journal of Interdisciplinary Research, 3*(5), 1891-1896.

Stern, H. H. (1983). *Fundamental concepts of language teaching*. Oxford: Oxford University Press.

Storch, S.A, & Whitehurst, G. J. (2002). Oral language and code-related precursors to reading: Evidence from a longitudinal structural model. *Developmental Psychology, 38*, 934–947.

Tamis-LeMonda, C. S., & Rodriguez, E. T. (2009). *Parents' role in fostering young children's learning and language development.* Language Development and Literacy: Encyclopedia on Early Childhood Development. Retrieved from http://www.child-encyclopedia.com/sites/default/files/textes-experts/en/622/parents-role-in-fostering-young-childrens-learning-and-language-development.pdf.

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education, 2*, 53-55.

Widdowson, H. G. (1978). *Teaching language as communication.* Oxford: Oxford University Press.

Xue-Ping, G. (1997). A scheme for the obtaining of language skills. *The Internet TESL Journal, 3*(6).

van der Slik, F. W. P., van Hout, Roeland, W. N. M., & Schepens, J. J. (2015). The gender gap in second language acquisition: Gender differences in the acquisition of Dutch among immigrants from 88 countries with 49 mother tongues. *PLoS One, 10*(11).