Ahi Evran Üniversitesi
Kırşehir Eğitim Fakültesi Dergisi

ISSN: 2147 - 1037

http://kefad.ahievran.edu.tr

Investigation of Gender Bias in TIMSS 2015 and 2019 Mathematics Items in Turkey: Differential Item Functioning Analysis with the SIBTEST Procedure

Musa Sadak

Article Information

CrossMark

DOI: 10.29299/kefad.961858

Received: 03.07.2021
Revised: 26.02.2022
Accepted: 17.03.2022

Keywords:

Item bias,
Gender equity,
Differential Item Functioning (DIF) analysis,
SIBTEST procedure,
Mathematics education,
TIMSS

Abstract

The aim of this study is to determine whether eighth- grade mathematics items in TIMSS international assessment (*f*=483), 224 in the TIMSS 2015, and 259 in the TIMSS 2019, have item bias between male and female students in Turkey and if any, to carry out detailed investigations based on TIMSS conceptual framework. The sample of the study consists of a total of 5,080 students, including 3,058 8th graders (n_{girls}= 1,577, n_{boys}= 1,481) in TIMSS 2015, and 2,022 8th graders (n_{girls}= 1,027, n_{boys}= 995) in TIMSS 2019. After the Confirmatory Factor Analysis was performed for the mathematics booklets used in two assessments as a prerequisite for the analysis of the data, items indicating bias were determined using the SIBTEST procedure, which is a special analysis method for the Differential Item Functioning (DIF) statistical technique. As a result of the analysis, it was concluded that 25 out of 224 mathematics items in TIMSS 2015 and 15 out of 259 mathematics items in TIMSS 2019 exam had item bias based on genders. The distribution of these biased items by gender and finally by content and cognitive domains within the TIMSS conceptual framework is presented. The findings were discussed in line with the literature and relevant suggestions were also provided.

TIMSS 2015 ve 2019 Matematik Sorularının Türkiye'de Cinsiyete Göre Madde Yanlılığının İncelenmesi: SIBTEST Prosedürü ile Değişen Madde Fonksiyonu Analizi

Makale Bilgileri

CrossMark

DOI: 10.29299/kefad.961858

Yükleme: 03.07.2021
Düzeltme: 26.02.2022
Kabul: 17.03.2022

Anahtar Kelimeler:

Madde yanlılığı,
Cinsiyet eşitliği
Değişen Madde Fonksiyonu (DMF) analizi,
SIBTEST prosedürü,
Matematik eğitimi,
TIMSS

Öz

Bu çalışmanın amacı TIMSS 2015 uluslararası sınavında 224 adet, TIMSS 2019'da ise 259 adet olmak üzere toplam 483 adet 8. sınıf matematik sorusunun Türkiye'de kız ve erkek öğrenciler arasında madde yanlılığına sahip olup olmadıklarını tespit etmek, varsa madde yanlılığı içeren soruların TIMSS kavramsal çerçevesi açısından ayrıntılı incelemelerini gerçekleştirmektir. Çalışmanın örneklemini TIMSS 2015'de 3,058 ($n_{kız}$=1,577, n_{erkek}=1,481) ve TIMSS 2019'da 2,022 ($n_{kız}$=1,027, n_{erkek}=995) olmak üzere toplam 5,080 öğrenci oluşturmaktadır. Verilerin analizi için öncelikle iki sınavda kullanılan matematik kitapçıkları için Doğrulayıcı Faktör Analizi gerçekleştirildikten sonra, Değişen Madde Fonksiyonu (DMF) için özel bir yöntem olan SIBTEST prosedürü kullanılmıştır. Sonuç olarak, TIMSS 2015'de toplam 224 matematik sorusunun 25'inin, TIMSS 2019'da ise, 259 matematik sorusunun 15'inin madde yanlılığına sahip olduğu sonucuna ulaşılmıştır. Bu soruların cinsiyetlere göre ve nihayetinde de TIMSS kavramsal çerçevesinde yer alan içerik ve bilişsel alanlara göre dağılımları sunulmuştur. Bulgular alan yazın doğrultusunda tartışılmış ve buna bağlı öneriler de sunulmuştur.

Sorumlu Yazar: Musa Sadak, Öğr. Gör. Dr., Kastamonu Üniversitesi, Türkiye, msadak@kastamonu.edu.tr, ORCID ID: 0000-0001-6036-1279.

Atıf için: Sadak, M. (2022). TIMSS 2015 ve 2019 matematik sorularinin Türkiye'de cinsiyete göre madde yanliliğinin incelenmesi: sibtest prosedürü ile değişen madde fonksiyonu analizi. *Kırşehir Eğitim Fakültesi Dergisi, 23*(Özel Sayı), 211-258.

Giriş

Uluslararası Eğitim Başarısını Değerlendirme Birliği (International Association for the Evaluation of Educational Achievement [IEA]) tarafından oluşturulan Uluslararası Matematik ve Fen Eğilimleri Çalışması (Trends in International Mathematics and Science Study [TIMSS]) ve Ekonomik İşbirliği ve Kalkınma Örgütü (Organization for Economic Co-operation and Development [OECD]) tarafından oluşturulan Uluslararası Öğrenci Değerlendirme Programı (Programme for International Student Assessment [PISA]) gibi uluslararası değerlendirme sınavları ülkelere matematik ve fen alanındaki performanslarını diğer katılımcı ülkeler arasında kıyaslama fırsatı sunmaktadır (Akyüz, 2014; Akyüz ve Berberoğlu, 2010; Doğan ve Barış, 2010; İncikabı, 2012). Bu değerlendirmeler 1960'lardan beri kullanılmaktadır (Yıldırım, Yıldırım, Ceylan, Yetişir ve Ajans, 2013). Uluslararası eğitim değerlendirmelerinden elde edilen sonuçlar ülkelere, sonuçların nasıl yorumlandığına bağlı olarak öğrencilerin performanslarını iyileştirmek için eğitim politikalarını şekillendirmek amacıyla bir bakış açısı kazandırmaktadır (Akyüz, 2006; Akyüz, 2014; Akyüz ve Berberoğlu, 2010; Bilican, Demirtaşlı ve Kilmen, 2011; Doğan ve Barış, 2010; İncikabı, 2012). Bu nedenle ülkeler, öğrencilerinin uluslararası ortamdaki başarılarının ilerlemesini izlemek ve başarılarını etkileyen faktörleri araştırmak için bu uluslararası değerlendirmelere katılırlar (Akyüz, 2014; Doğan ve Barış, 2010; İncikabı, 2012). Bu faktörleri anlayarak ve eğitim sistemlerini diğer eğitim sistemleriyle karşılaştırarak, eğitim politika yapıcıları kararlarını değerlendirebilir, sorunları belirleyebilir ve daha etkili politikalar geliştirebilirler (Akyüz ve Berberoğlu, 2010; Bilican ve diğerleri, 2011).

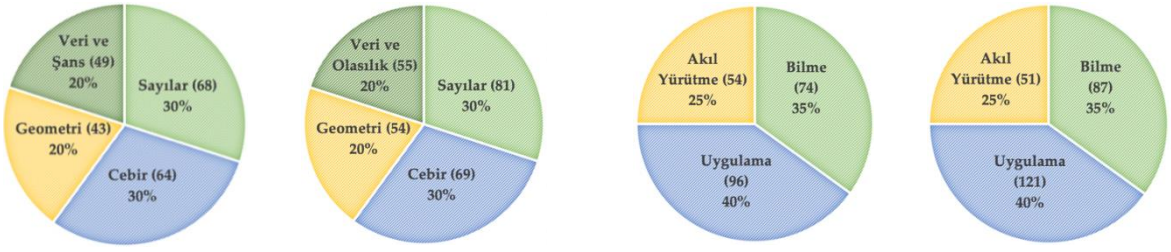
TIMSS, dünyanın dört bir yanındaki katılımcı ülkelerdeki 4. ve 8. sınıf öğrencilerinin matematik ve fen bilgisi alanlarındaki performanslarını değerlendirmek için her dört yılda bir tekrarlanan uluslararası bir değerlendirmedir (Doğan ve Barış, 2010; İncikabı, 2012). Türkiye bu değerlendirmeye ilk olarak 1999 yılında katılmıştır. 1999 yılındaki değerlendirmeye sadece sekizinci sınıf düzeyinde katılırken, 2003 yılında katılım göstermemiş, 2007 yılında ise yine sekizinci sınıf düzeyinde katılmıştır. 2011 yılına gelindiğinde, Türkiye artık hem dördüncü hem de sekizinci sınıf düzeyinde katılım göstermeye başlamış (Bilican ve diğerleri, 2011; Erkan, 2013; Güner, Sezer ve İspir, 2013) ve bu durum 2015 (Mullis, Martin, Foy ve Hooper, 2016) ve 2019 yıllarında da (Mullis, Martin, Foy, Kelly ve Fishbein, 2020) devam etmiştir. 2019 yılında uygulanan sınavda diğerlerinden farklı olarak Türkiye elektronik olarak da bu sınava katılmıştır. Katıldığı yıllar içerisinde diğer katılımcılara göre Türkiye'nin konumu Tablo 1'de verilmiştir.

Tablo 1. *Türkiye'nin katıldığı TIMSS sınavları ve katılımcılar arasındaki konumu*

	4. sınıf		8. sınıf	
	Türkiye Sıralama	Toplam Katılımcı	Türkiye Sıralama	Toplam Katılımcı
TIMSS 1999 (Akyüz, 2006)	-	-	31	38
TIMSS 2007 (Martin, Mullis ve Foy, 2008)	-	-	37	67 (8)
TIMSS 2011 (Mullis, Martin, Foy ve Arora, 2012)	40	60 (8)	37	59 (14)
TIMSS 2015 (Mullis ve diğerleri, 2016)	41	57 (7)	29	46 (7)
TIMSS 2019 (Mullis ve diğerleri, 2020)	26	64 (6)	24	46 (7)

Not: Parantez içerisinde verilen sayılar toplam katılımcılar arasında bazı ülkelerden özel olarak katılım gösteren bölgeleri belirtmektedir (*benchmarking participants*).

Tablo 1'de görüldüğü üzere, Türkiye yıllar içerisinde nispeten konumunu geliştirmiştir. Özellikle 2015 ve 2019 yıllarındaki performansı ile Türkiye diğer yıllara nazaran bir gelişim eğilimine girmiştir. Özellikle bu iki yıldaki değişimleri anlamlı olarak okuyabilmek sonraki yıllarda da artışı devam ettirmek adına önemlidir. Bu anlamda, öğrencilerin diğer ülkeler nezdindeki başarılarını ve bu başarı veya başarısızlığın arkasında yatan sebepleri ortaya koymak adına düzenlenen TIMSS uluslararası sınavının matematik bölümünde, öğrencilerin hem içerik hem de bilişsel anlamda gelişimleri gözlemlenebilmektedir. Bu sınavda sorular belli içerik ve bilişsel alanlar etrafında hedeflenen doğrultuda oluşturulmuştur. 8. sınıf sorularının içerdikleri içerik alanları TIMSS 2015 sınavında *sayılar*, *cebir*, *geometri* ve *veri ve şans* iken TIMSS 2019 sınavında ise veri ve şans alt içerik alanı *veri ve olasılık* olarak değişmiştir. Diğer yandan bilişsel alanlar ise her iki yılda da *bilme*, *uygulama* ve *akıl yürütme* olarak belirtilmiştir (Gronmo, Linquist, Arora ve Mullis, 2013; Martin, Mullis ve Foy, 2017). Bu içerik ve bilişsel alanların hedeflenen yüzdelik dağılımları ile TIMSS 2015 ve TIMSS 2019 sınavlarındaki gerçek soru sayıları (parantez içerisinde) aşağıda verilmiştir (Şekil 1). TIMSS 2015 sınavında toplam 224 adet, TIMSS 2019 sınavında ise 259 adet 8. sınıf matematik sorusu bulunmaktadır. TIMSS 2019 sınavındaki sorular yeni oluşturulan elektronik sınavda kullanılan sorulardır.



TIMSS 2015 İçerik Alanlar | **TIMSS 2019 İçerik Alanlar** | **TIMSS 2015 Bilişsel Alanlar** | **TIMSS 2019 Bilişsel Alanlar**

Şekil 1. TIMSS 2015 ve TIMSS 2019 sınavlarında 8. sınıf düzeyinde kullanılan matematik sorularının içerik ve bilişsel alanlar yönünden hedeflenen (%) ve gerçek (parantez içinde) dağılımları (Gronmo ve diğerleri, 2013 ve Martin ve diğerleri, 2017'den derlenmiştir)

Cinsiyet ve Matematik Başarısı

Öğrencilerin başarısını etkileyen faktörler araştırmacıların yoğun ilgi gösterdiği konulardan biridir. Özellikle öğrenci ve aile ile ilgili faktörler (bkz. Akyüz, 2014; Atar, 2011; Demir, Kılıç ve Ünal, 2010), okulla ilgili faktörler (bkz. Demir ve diğerleri, 2010; Engin-Demir, 2009; Kılıç, Çene ve Demir,

2012), ve öğretmenlerle ile ilgili faktörler (bkz. Aaronson, Barrow ve Sander, 2007; Clotfelter, Ladd ve Vigdor, 2007; Clotfelter, Ladd ve Vigdor, 2010; Hill, Rowan ve Ball, 2005; Nye, Konstantopoulos ve Hedges, 2004; Stronge, Ward ve Grant, 2011) araştırmacıların ilgisini çekmiştir. Bu faktörler arasında öğrencilerin cinsiyetleri de başarıya etki eden önemli bir faktör olarak alan yazında yerini almıştır. Özellikle öğrenci başarısını modellemek adına yürütülen çalışmalarda, cinsiyet faktörü kontrol değişkeni olarak ele alınmış ve öğrenci başarısına etki eden faktörleri incelerken başarıya olan etkisi kontrol altına alınmak suretiyle çalışmalara dâhil edilmiştir (bkz. Aaronson ve diğerleri, 2007; Boyd, Grossman, Lankford, Loeb ve Wyckoff, 2005; Clotfelter ve diğerleri, 2010; Goldhaber ve Anthony, 2007; Hill ve diğerleri, 2005; Jacob ve Lefgren, 2002; Rowan, Chiang ve Miller, 1997; Stronge ve diğerleri, 2011). Türkiye'de yürütülen çalışmalarda ise cinsiyet, matematik başarısı açısından bir ikilem oluşturmaktadır. Bazı araştırmacılar öğrencilerin cinsiyetinin matematik başarısında önemli rol oynayan bir faktör olduğunu tespit ederken (Alacacı ve Erbaş, 2010; Demir ve diğerleri, 2010; Dinçer ve Kolasin, 2009; Gürsakal, 2012; Kılıç ve diğerleri, 2012), bazı araştırmacılar cinsiyet ile matematik başarısı arasında bir bağlantı olmadığını gösteren zıt sonuçları belirtmişlerdir (Aksu, 2001; Atar, 2011; Işıksal ve Aşkar, 2005). Çalışmalar genel itibariyle cinsiyeti öğrenci başarısına etki eden bir faktör olarak değerlendirmiş ve diğer değişkenlerle birlikte cinsiyet açısından da öğrenci başarısı açıklanmaya çalışılmıştır.

Yukarıda belirtilen içerik ve bilişsel alanlara göre cinsiyetler arasında matematik performansı açısından farklılıkları görebilmek adına, TIMSS 2015 ve TIMSS 2019 verileri üzerinden cinsiyetlere göre katılan bütün ülkelerin ortalama başarı puanları hem içerik hem de bilişsel alanlar yönünden aşağıda belirtilmiştir (Tablo 2). Tablo 2'de aynı zamanda Türkiye için de bu yıllardaki ortalama puanlar verilmiştir.

Tablo 2. *TIMSS 2015 ve TIMSS 2019 sınavlarında 8. sınıf düzeyinde cinsiyet açısından içerik ve bilişsel alanlarda Türkiye ve TIMSS genel ortalama başarı puanları*

		Türkiye 2015		TIMSS 2015		Türkiye 2019		TIMSS 2019	
		Kız	Erkek	Kız	Erkek	Kız	Erkek	Kız	Erkek
İçerik Alanları	Sayılar	461	455	484	491*	496	490	493	497*
	Cebir	443	452*	492*	481	503*	482	503*	493
	Geometri	469*	450	487*	480	496*	483	499*	495
	Veri ve Olasılık	472*	454	481*	479	506	498	490	489
Bilişsel Alanlar	Bilme	470	464	488*	485	503*	485	499*	494
	Uygulama	450	444	486	486	494	488	497	496
	Akıl Yürütme	461	458	487*	482	511*	497	501*	497
	Genel Ortalama	461	455	488*	485	501	490	491	488

*Anlamlı ölçüde diğer cinsiyet değerinden yüksek (.05 seviyesinde)

(National Center for Education Statistics, 2021 ve Mullis ve diğerleri, 2020'den faydalanılarak derlenmiştir)

Tablo 2'de görüldüğü üzere, sınav genel ortalama puanlarına baktığımızda kız ve erkek öğrenciler arasında ne TIMSS ortalaması bakımından ne de Türkiye bakımından istatistiksel olarak anlamlı ölçüde bir fark gözlemlenmemiştir (TIMSS 2015 genel ortalama hariç). İçerik ve bilişsel alanlar açısından baktığımızda ise, TIMSS sınavına katılan tüm ülkelerin ortalamaları üzerinden kız

öğrencilerin içerik alanları ve bilişsel alanlarda erkek öğrencilere göre nispeten daha yüksek ortalama puana sahip oldukları gözlemlenmiştir. Katılımcı tüm ülkelerin ortalamalarına bakıldığında, 2015 ve 2019 yıllarında erkek öğrenciler dört içerik alanından sadece sayılar içerik alanında kız öğrencilerden ortalama olarak anlamlı ölçüde daha başarılı olmuşlardır. Bilişsel alanlar açısından ise kız öğrencilerin bilme ve akıl yürütme bilişsel alanlarında erkek öğrencilerden istatistiksel olarak anlamlı ölçüde daha başarılı olduğu, erkek öğrencilerin ise uygulama içerik alanında nispeten kız öğrencilere göre daha başarılı oldukları gözlemlenmiştir.

Türkiye açısından Tablo 2 incelendiğinde ise, 2015 yılında içerik alanları açısından kız öğrenciler geometri ve veri ve olasılık içerik alanlarında erkek öğrencilerden ortalama olarak anlamlı ölçüde daha yüksek puana sahip olmakla birlikte, erkek öğrenciler ise cebir alt içerik alanında kız öğrencilerden daha yüksek ortalama puana sahiptirler. 2015 yılında bilişsel alanlarda Türkiye'deki kız ve erkek öğrenciler arasında anlamlı bir farklılık gözlemlenmemiştir. 2019 yılına geldiğimizde ise, Türkiye'de TIMSS'e katılım gösteren ülkelerin ortalamalarına benzer şekilde kız öğrencilerin içerik ve bilişsel alanlarda erkek öğrencilere göre daha yüksek ortalama puana sahip oldukları görülmektedir. Erkek öğrenciler hiçbir içerik ve bilişsel alanda kız öğrencilerden anlamlı ölçüde yüksek puana sahip olamamışlardır. Genel ortalamalar açısından bu iki cinsiyet arasında ne Türkiye ne de ülkeler genel ortalamaları açısından anlamlı ölçüde fark gözlemlenmezken, içerik ve bilişsel alanlarda özellikle kız öğrenciler tarafına olumlu yönde bir eğilim bulunmaktadır. Bu anlamda cinsiyetler arasında başarı farklılıklarını içerik ve bilişsel alanlar yönünden okumak önem arz etmektedir.

Önceki araştırmalara bakıldığında, TIMSS ve PISA sınavlarına katılan 69 ülke üzerinden yapılan bir çalışmaya göre her ne kadar başarı anlamında aralarında çok farklar olmasa da erkek öğrencilerin matematiğe karşı kız öğrencilere nispeten daha olumlu tutumlar sergilediği görülmektedir (Else-Quest, Hyde ve Linn, 2010). Peki bu tutum başarılarına da yansımış mıdır? Bu açıdan, matematik içerik ve bilişsel alanlarında yapılan çalışmalara da değinmek gerekir. İçerik alanlar açısından bakıldığında, erkek öğrencilerin geometri alanında kız öğrencilere nispeten daha başarılı olduklarını, kız öğrencilerin ise cebir alanında erkek öğrencilere nazaran daha başarılı olduklarını belirten araştırmalar bulunmaktadır (Lane, Wang ve Magone, 1996; McGraw, Lubienski ve Strutchens, 2006). Bilişsel alanlar açısından ise, akıl yürütme problemlerinde kız öğrencilerin erkek öğrencilere nispeten daha başarılı oldukları önceki araştırmacılar tarafından ortaya konulmuştur (Friedman, 1996; Ryan ve Chiu, 1996).

Türkiye örneklemi kullanılarak yapılan çalışmalara da değinmek gerekir. Cinsiyet ve matematik başarısı arasındaki ilişkiyi inceleyen çalışmalardan özellikle uluslararası sınavların verilerini kullananlara bakacak olursak, Alacacı ve Erbaş'ın (2010) PISA 2006 sonuçları üzerinde Hiyerarşik Lineer Modelleme (HLM) kullanarak yaptıkları araştırmaya göre cinsiyetin başarı üzerinde önemli bir gösterge olduğu bulunmuştur. Elde ettikleri sonuçlar erkek öğrencilerin kız öğrencilere göre daha

başarılı olduğunu göstermektedir. Demir ve diğerlerinin (2010) yine HLM kullanarak PISA 2006 sonuçları üzerine yaptığı çalışmada, erkek öğrencilerin matematikte daha iyi puanlara sahip oldukları bulgusu desteklenmektedir. Dinçer ve Kolasin (2009) PISA 2006'daki matematik testinde kız öğrencilerin erkek öğrencilere göre 14 puan daha düşük ortalama puana sahip olduklarını; ancak okuma testinde erkek öğrencilere göre 32 puan daha yüksek bir ortalamaya sahip olduklarını belirtmişlerdir. PISA 2009'da, öğrencilerin cinsiyetiyle ilgili sonuçlar Türkiye için yine benzerlik göstermiştir. Öğrencilerin cinsiyeti, öğrencilerin matematik başarısı üzerinde anlamlı bir etkiye sahiptir (Kılıç ve diğerleri, 2012). Başarı durumlarına ek olarak, PISA 2009 sınavında beklenti yönünden de erkek öğrencilerin kız öğrencilere göre daha yüksek performans gösterme beklentisi içinde oldukları belirtilmiştir (Gürsakal, 2012).

Ancak, TIMSS 2007'de Madde Tepki Kuramı (Item Response Theory – IRT) temelinde oluşturulan modellemeye göre cinsiyet değişkeninin Türkiye'de öğrencilerin matematik başarısının nötr bir göstergesi olduğu bulunmuştur (Atar, 2011). Yine Aksu'nun (2001) Ankara'da özel bir okulda öğrenim gören öğrenciler üzerine gerçekleştirdiği çalışmaya göre, öğrencilerin cinsiyetleri ile performansları arasında anlamlı bir ilişki bulunamamıştır. Ek olarak, Işıksal ve Aşkar'ın (2005) Türkiye'de yedinci sınıfta bulunan 64 öğrenci üzerinden yaptıkları çalışmada, cinsiyetler arasında ne matematik performansları ne de matematiksel özgüven anlamında ortalama puan yönünden anlamlı bir farklılık olmadığını belirtmişlerdir.

Özetle, Türkiye'de özellikle PISA sınavında erkek öğrencilerin daha başarılı olduğu sonuçlara nispeten (Alacacı ve Erbaş, 2010; Demir ve diğerleri, 2010; Dinçer ve Kolasın, 2009; Kılıç ve diğerleri, 2012; Gürsakal, 2012), bazı çalışmalarda ise cinsiyet ve başarı arasında anlamlı bir bağlantının tespit edilmediği görülmektedir (Aksu, 2001; Atar, 2011; Işıksal ve Aşkar, 2005). Diğer bir yandan, TIMSS 2015 ve TIMSS 2019 sonuçlarına genel ortalama puan yönünden bakıldığında kız ve erkek öğrenciler arasında anlamlı bir farklılık gözlemlenmemiştir. Ancak katılımcı ülkeler genelinde veya Türkiye özelinde kız öğrencilerin içerik ve bilişsel alanlar için hesaplanan ortalama puanlar açısından erkek öğrencilere göre nispeten anlamlı ölçüde daha yüksek puanlara sahip oldukları görülmektedir. Önceki araştırmalara göre ise erkek öğrencilerin geometri, kız öğrencilerin ise cebir içerik alanında daha başarılı oldukları gözlemlenmiştir (Lane ve diğerleri, 1996; McGraw ve diğerleri, 2006). Ek olarak, kız öğrencilerin akıl yürütme bilişsel alanında erkek öğrencilere göre daha başarılı oldukları da önceki araştırmacılar tarafından belirtilmiştir (Friedman, 1996; Ryan ve Chiu, 1996).

Sınavlarda Cinsiyet Yanlılıkları

Kız ve erkek öğrencilerin matematik başarıları arasındaki farkı anlamanın yollarından biri de sınavlarda kullanılan soruların yapısını incelemek ve olası yanlılıkları belirleyebilmektir. Yukarıda da değinildiği üzere, bu iki grup öğrencinin farklı ölçme araçlarıyla ölçülen matematik performansları farklılık göstermektedir. Bu iki cinsiyet arasında oluşan performans farklılıklarının sınavlarda

kullanılan soruların yapısından kaynaklanma ihtimali göz ardı edilmemelidir. Bakan-Kalaycıoğlu ve Kelecioğlu'nun (2011), Türkiye'de 2005 yılında uygulanan Öğrenci Seçme Sınavı (ÖSS) üzerinde cinsiyetler arasında yanlılık gösteren soruları belirlemek adına yaptıkları Değişen Madde Fonksiyonu (DMF) analizine göre, 45 sorudan oluşan matematik alt testinde biri cebir ikisi de geometri alt alanlarında olmak üzere üç adet soruda DMF tespit etmişlerdir. DMF tespit edilmesi, soruların gruplardan birine yanlılık gösterebilmesi ihtimalini işaret etmektedir. Burada cebir sorusu kız öğrenciler lehine DMF belirtirken, iki geometri sorusu ise erkek öğrenciler lehine DMF göstermiştir. DMF analizi bu çalışmada da kullanılan bir yöntem olup, önyargı (veya yanlılık) belirtebilecek soruları tespit etmek için kullanılan özel bir istatistiksel yöntemdir. Çalışmanın yöntem bölümünde bu analiz tekniği ile ilgili ayrıntılı bilgi sağlanmıştır.

Ek olarak, Karakaya (2012), 2009 yılında uygulanan ve Türkiye'de 8. sınıf öğrencilerinin liselere girebilmek adına katıldıkları Seviye Belirleme Sınavındaki (SBS) matematik soruları üzerine yaptıkları DMF analizine göre iki soruda DMF tespit etmişler, ancak uzman görüşü neticesinde bu soruların cinsiyetler lehine yanlılık belirtmediğini ortaya koymuşlardır. Belirtilen çalışmada, Klasik Test Kuramına (KTK – *Classical Test Theory*) göre oluşturulmuş olan Mantel-Haenszel metodu tercih edilmiştir. Yine 2009 yılındaki SBS sınavı için uygulanan başka bir DMF çalışmasında, üç farklı DMF analiz metodu ile cinsiyet ve okul türleri açısından matematik sorularının yanlılık belirtip belirtmediği incelenmiştir (Kelecioğlu, Karabay ve Karabay, 2014). Belirtilen çalışmada ise 20 matematik sorusundan 14'ünde DMF belirtildiği sonucuna varılmış, ancak bunların 3'ünün uzman görüşü neticesinde yanlılık teşkil etmediği sonucuna ulaşılmıştır. Buradan hareketle, aynı sınavın soruları üzerinde yapılan aynı analiz içerisinde kullanılan yöntemin önemi ortaya çıkmaktadır. Kan, Sünbül ve Ömür'ün (2013) yine SBS sınavı üzerine ama 2011 yılında uygulanan formatıyla yaptıkları ve KTK'den ziyade Madde Tepki Kuramı (MTK – *Item Response Theory*) tabanlı DMF analizlerinde, Lord yöntemine göre 20 adet matematik sorusunun 20'sinin de ve Raju yöntemine göre ise 15'inde kız veya erkek öğrenciler lehine yanlılık tespit ettiklerini belirtmişlerdir. KTK, öğrencilerin bir testteki skorları veya tahmin edilen skorları üzerinde bir raporlama sistemi üzerine kurulu iken, MTK aksine öğrencilerin beceri seviyeleri ve testteki her bir sorunun karakteristik değerleri üzerinden bağlantılar kurma mantığı üzerine dayalıdır (Hambleton ve Jones, 1993). Hambleton ve Jones (1993) yine bu iki kuram için, KTK için gerekli varsayımların zayıf olduğunu yani bu kurama yönelik analizler yapabilmek için gerekli olan varsayımların sağlanmasının daha kolay olduğunu, ancak MTK için gerekli varsayımların çok daha güçlü olduğunu da belirtmişlerdir. Bu sebeple, MTK tabanlı bir DMF analizinin sorularda olabilecek yanlılıkları tespit ederken daha hassas davrandığı görülmektedir.

Sonuç olarak, Türkiye'de bir örneklem oluşturmak suretiyle yapılan deneysel çalışmalarda cinsiyetler arasında matematik performansı yönünden anlamlı bir farklılık gözlemlenmezken, özellikle ulusal ve uluslararası sınavlarda (özellikle TIMSS ve PISA) bu iki cinsiyet arasında matematik performansı yönünden farklılıklar ortaya konulmuştur. Bu anlamda, deneysel çalışmalarda

gözlemlenmeyen bu farklılıkların ulusal veya uluslararası düzeydeki çalışmalarda gözlemlenmesi, sınavlarda kullanılan soruların yapısı ile de ilgili olabilir. Diğer bir yandan, Türkiye'deki TIMSS ve PISA sonuçlarına göre erkek veya kız öğrencilerin daha başarılı olma durumları değişiklik göstermektedir. Bu anlamda, yukarıda ayrıntılı olarak sonuçlarına yer verilen ve soru yapıları itibariyle farklı cinsiyetlerde yanlılık olma durumlarını inceleyen çalışmalara göre, özellikle Türkiye'de yerleştirme amaçlı kullanılan sınavlarda matematik sorularının yanlılık gösterdiği görülmektedir. Haliyle, ulusal sınavlar üzerine yapılan yanlılık çalışmalarında Türkiye'de cinsiyetler arasındaki performans farklılıklarının soru yapılarından kaynaklı olabilecek kısımları araştırmacılar tarafından incelenmiştir. Ancak, Türkiye'deki öğrenciler üzerinden uluslararası sınav verileri kullanılarak bu yanlılıkları ortaya koyan çalışmalara yeterince rastlanılmamaktadır. Özellikle de, sınavlarda kullanılan soruların yine sınavlarda belirtilen içerik ve bilişsel alanlar yönünden yanlılıklarını ortaya koyan bir çalışmaya ihtiyaç duyulduğu öğrenciler arasında bu alanlardaki başarı farklılıkları da göz önünde bulundurulduğunda (Friedman, 1996; Lane ve diğerleri, 1996; McGraw ve diğerleri, 2006; Ryan ve Chiu, 1996) daha net olarak ortaya çıkmaktadır.

Bu çalışmada Türkiye'de TIMSS 2015 ve TIMSS 2019 uluslararası sınavlarında kullanılmış olan 8. sınıf matematik sorularının kız ve erkek öğrenciler açısından madde yanlılıklarını tespit etmek amaçlanmıştır. Diğer bir ifadeyle, kullanılan soruların yapı itibariyle farklı cinsiyetlerdeki öğrenciler açısından avantaj veya dezavantaj oluşturma durumları ortaya konulmak istenmiştir. Her ne kadar bu iki grup öğrencinin teorik olarak sınav sorularının ölçmek istediği yapılar üzerinde eşit düzeyde bilgi düzeyine sahip oldukları kabul edilse de soruların yapısı itibariyle cevaplarda farklılıklar gözlenebilmektedir. Özellikle kız ve erkek öğrencilerin matematik dersi konusundaki yaklaşımları ve sorularda kullanılan ifadelerin bu farklı yaklaşımlara yakınlığının farklı olabilmesinden kaynaklı farklılıklar bu iki cinsiyet arasında soruların farklı şekilde davranmasında rol oynayabilmektedir. Bu nedenle, bu çalışma, Türkiye'deki kız ve erkek öğrenci grupları üzerinde farklı şekilde çalışan TIMSS 2015 ve TIMSS 2019 8. sınıf matematik sorularını özel bir analiz yöntemi olan Değişen Madde Fonksiyonu (DMF) (*Differential Item Functioning – DIF*) ile ortaya koymayı amaçlamaktadır. Bu analiz sonucunda, kız ve erkek öğrenciler açısından yanlı olarak işlev gören matematik soruları tespit edilmeye ve bu soruların TIMSS sınavında belirlenmiş olan içerik ve bilişsel alanlar açısından fark edilebilir bir özellik gösterip göstermedikleri ortaya konulmak istenmiştir. Yanlılık belirten soruların sadece cinsiyet farklılığından dolayı yanlı işlev gördüklerini söylemek mümkün olmasa da bu soruların oluşturdukları içerik ve bilişsel alanları cinsiyet açısından daha detaylı incelemek adına yardımcı olacaklardır. Çalışma, aşağıdaki araştırma soruları özelinde şekillenmiştir:

- TIMSS 2015 ve TIMSS 2019 uluslararası sınavlarında cevaplanan 8. sınıf matematik soruları Türkiye'deki kız ve erkek öğrenciler arasında yanlılık oluşturmakta mıdır?

– Yanlılık gösterdiği tespit edilen sorular varsa, bu sorular yine TIMSS 2015 ve TIMSS 2019 kavramsal çerçevelerinde belirtilen içerik alanları ve bilişsel alanlar açısından nasıl bir özellik göstermektedir?

Yöntem

Veri Toplama

Çalışmanın örneklemini Türkiye'de 8. sınıf düzeyinde TIMSS 2015 ($n_{kız}$= 1,577, n_{erkek}= 1,481, n_{toplam}= 3,058) ve TIMSS 2019 ($n_{kız}$ = 1,027, n_{erkek} = 995, n_{toplam} = 2,022) sınavlarının matematik bölümüne katılan ve tek sayılı soru kitapçıklarını cevaplayan toplam 5,080 öğrenci oluşturmaktadır. Bu öğrencilerin TIMSS 2015 için 224, TIMSS 2019 için de 259 olmak uzere toplam 438 farklı matematik sorusuna verdikleri cevaplar üzerinden kız veya erkek öğrenci grupları için yanlılık gösteren sorular tespit edilmeye çalışılmıştır.

TIMSS 2015 ve TIMSS 2019 sınavına katılan öğrenciler 14 farklı kitapçıktan birini yanıtlamışlardır. Her bir kitapçıkta yine 14 farklı matematik soru bloklarından ikisi bulunmaktadır. Öğrencilerin almış oldukları kitapçıklara göre her iki sınavda da cevapladıkları matematik soru blokları ve bu bloklardaki soru sayıları aşağıda gösterilmiştir (Tablo 3).

Tablo 3. *Matematik blokları ve soru sayıları açısından TIMSS 2015 ve TIMSS 2019 8. sınıf soru kitapçıkları*

	Soru Blokları		**TIMSS 2015**			**TIMSS 2019**		
	Blok-1	Blok-2	Blok-1	Blok-2	Toplam	Blok-1	Blok-2	Toplam
Kitapçık 1	M01	M02	17	18	35	16	24	40
Kitapçık 2	M02	M03	18	15	33	24	16	40
Kitapçık 3	M03	M04	15	18	33	16	27	43
Kitapçık 4	M04	M05	18	19	37	27	18	45
Kitapçık 5	M05	M06	19	15	34	18	14	32
Kitapçık 6	M06	M07	15	17	32	14	16	30
Kitapçık 7	M07	M08	17	15	32	16	21	37
Kitapçık 8	M08	M09	15	15	30	21	18	39
Kitapçık 9	M09	M10	15	14	29	18	18	36
Kitapçık 10	M10	M11	14	15	29	18	15	33
Kitapçık 11	M11	M12	15	14	29	15	16	31
Kitapçık 12	M12	M13	14	16	30	16	16	32
Kitapçık 13	M13	M14	16	16	32	16	25	41
Kitapçık 14	M14	M01	16	17	33	25	16	41

(Martin, Mullis ve Foy, 2013 ve Martin ve diğerleri, 2017'den derlenmiştir)

Tablo 3'te gösterildiği üzere, tek sayıda kitapçıkları alan öğrenciler incelendiğinde olası bütün matematik soru bloklarının incelenebilmesi mümkündür. M01'den başlamak üzere M14'e kadar toplam 14 farklı blok halinde matematik soruları bulunmaktadır. Hem TIMSS 2015 hem de TIMSS 2019 sınavları için tek veya çift sayılı kitapçıklar kullanılırsa olası bütün soruları madde yanlılığı açısından incelemek mümkün olacaktır. Bu yüzden, fark olmamakla birlikte tek sayılı kitapçıkların kullanılması tercih edilmiştir. Tablo 4 tek sayılı olan her bir kitapçığı alan öğrenci sayılarını ve cinsiyet dağılımlarını göstermektedir.

Tablo 4. *Türkiye'deki belirlenen TIMSS 2015 ve TIMSS 2019 matematik soru kitapçıkları ve cevaplayan 8. sınıf öğrencilerinin cinsiyet olarak dağılımı (çalışmanın örneklemi)*

		Soru Sayısı		Öğrenci Sayısı					
		TIMSS	TIMSS	TIMSS 2015			TIMSS 2019		
Kitapçık ve Soru Blokları		2015	2019	Kız	Erkek	Toplam	Kız	Erkek	Toplam
Kitapçık 1	M01-M02	35	40	228	207	435	157	129	286
Kitapçık 3	M03-M04	33	43	229	211	440	144	143	287
Kitapçık 5	M05-M06	34	32	229	211	440	139	149	288
Kitapçık 7	M07-M08	32	37	221	211	432	152	135	287
Kitapçık 9	M09-M10	29	36	218	217	435	140	148	288
Kitapçık 11	M11-M12	29	32	218	223	441	147	142	289
Kitapçık 13	M13-M14	32	39	234	201	435	148	149	297
	Toplam	224	259	1,577	1,481	3,058	1,027	995	2,022

Veri Analizi

Çalışma doğası itibariyle nicel bir çalışma olup, maddelerin cinsiyetlere göre yanlılıklarını belirlemek üzere Değişen Madde Fonksiyonu (DMF) (Differential Item Functioning – DIF) tekniğinden faydalanılmıştır. DMF gerçekleştirilmeden önce her bir gruptaki matematik sorularının tek boyutlu yapıda olması gereklidir (Jöreskog ve Sörbom, 1996). Bu nedenle ilk olarak, R Studio açık kaynak istatistik programı (R Core Team, 2018) üzerinden *lavaan* (Rosseel, 2012) paketi kullanılarak Doğrulayıcı Faktör Analizi (Confirmatory Factor Analysis – CFA) gerçekleştirilmiştir. Soruların tek boyutlu olması şartını her bir kitapçık için inceledikten sonra da yine R Studio programı üzerinden *difR* paketi (Magis, Beland, Tuerlinckx ve De Boeck, 2010) içerisindeki *difSIBTEST* (Shealy ve Stout, 1993) modülü kullanılarak SIBTEST prosedürüne uygun olacak şekilde DMF analizi gerçekleştirilmiştir.

Madde yanlılığı, farklı gruplarda yer alan öğrenciler soruların ölçmek istediği örtük değişken üzerinde teorik olarak eşit beceriye sahip olsalar bile farklı başarı performansları göstermeleri olasılığıdır (Zumbo, 1999). Diğer bir ifadeyle, her ne kadar Türkiye'deki kız ve erkek öğrenciler sınavlarda eşit düzeyde başarı beklentisine sahip olsalar da bazı soruların bu cinsiyetlerden birine yanlı davranması sorunun yapısı itibariyle mümkün olabilir. DMF analizi tekniği sayesinde kız ve erkek öğrenciler üzerinde yanlı davranan matematik sorularını tespit etmek mümkündür. Tespit edilen soruların da içerik ve bilişsel alanlar yönünden incelenmesi önemlidir.

Değişen Madde Fonksiyonu (DMF) (Differential Item Functioning – DIF) Analizi: Madde Tepki Kuramı (Item Response Theory – IRT) dünyasında DMF, madde veya test sapması kavramlarının yerini alan bir analiz yöntemidir (bkz. Embretson ve Reise, 2013; Zumbo, 1999). DMF, yanlılık adayı olan bir maddenin, o maddenin ölçmek istediği örtük değişkenle aynı ilişkiye sahip olma açısından gruplar arasında farklılık oluştuğunda ortaya çıkar (bkz. Embretson ve Reise, 2013). Ayrıca, Embretson ve Reise (2013) DMF analizindeki önemli bir gelişme olarak Shealy ve Stout'un (1993) DMF için oluşturdukları çok boyutlu SIBTEST (Simultaneous Item Bias Test) modelini belirtmişlerdir. Shealy ve Stout (1993), DMF analizinde karşılaştırılan grupları referans ve odak grupları olarak tanımlamıştır. Diğer DMF belirleme yöntemlerine göre SIBTEST prosedürü daha yeni bir yöntem olarak ortaya çıkmakta ve çok

sık olarak kullanılan Mantel-Haenszel yöntemiyle elde edilen ortalama puan değerlerinin bireylerin testten genel olarak aldıkları toplam puan doğrultusunda doğrusal regresyon ile düzenlenmesi sistemine dayanmaktadır (Osterlind ve Everson, 2009). Shealy ve Stout (1993) oluşturdukları yöntem olan SIBTEST'in prosedürünü, SIBTEST'in birkaç maddeli bir testte DMF'yi inceleyebileceği ve maddelerdeki yanlılığı/DMF'yi bilişsel olarak inceleme fırsatı sağlayabileceğini belirtmektedirler. Ayrıca Awuor (2008), SIBTEST'in Shealy ve Stout'un (1993) çalışmasına bir uzantı olarak geliştirilen parametrik olmayan bir prosedür olduğunu ve tahmin edilen parametreler yerine gerçek madde yanıt verilerini kullandığını belirtmektedir. Shealy ve Stout (1993), SIBTEST'in birden fazla maddeye sahip bir test boyunca veya sadece bir madde için DMF'yi algılayabildiğini belirtmiştir. Embretson ve Reise (2013) , SIBTEST prosedürünü çok kulağa hoş bulmuş ve modelin arkasındaki teorinin gerçek dünya testlerine çok iyi uyduğunu ve yakın gelecekte araştırmacılardan daha fazla ilgi göreceğini belirtmiştir. Shealy ve Stout'un (1993) çalışmasının bir uzantısı olarak, Stout ve Roussos (1996) tarafından SIBTEST bilgisayar programı geliştirilmiştir. Bu bilgisayar programı, daha sonra R Studio istatistik programındaki DMF analizleri için geliştirilmiş kapsamlı bir paket olan difR paketi içerisinde difSIBTEST isimli bir modül olarak yerini almıştır. Bu çalışma için de bu modül kullanılmıştır.

Bu çalışmada referans grubu Türkiye'de TIMSS 2015 ve TIMSS 2019 uluslararası sınavlarına 8. sınıf matematik alanında katılan kız öğrenciler olmakla birlikte, odak grubu da yine aynı sınavlara katılan erkek öğrenciler olarak belirlenmiştir. Embretson ve Reise (2013) bu grupların (referans ve odak) farklı ortalamaya ve örtük değişken üzerinden standart sapmaya sahip olabileceklerini belirtmiştir. Ancak, bu farklılıkların DMF belirtileri olmadığını, araştırma ortamına bağlı olarak DMF analiz prosedürünü karmaşıklaştırabileceklerini belirtmişlerdir. Ama Zumbo (1999), maddeye doğru cevap verme olasılıklarının belirlenmesi açısından DMF analizinin uygulanabilmesi için bu iki grubun ilgili özellik değişkeni üzerinde eşleştirilmesi gerektiğine işaret etmiştir. Bir örtük özellik değişkeni üzerinde DMF analizini uygularken bir diğer önemli husus, DMF'yi belirlemekle ilgilenen araştırmacıların referans ve odak grupları için aynı öğeleri kullanmasıdır (bkz. Embretson ve Reise, 2013).

DMF analizi için kabul edilen boş hipotez (null hypothesis – H_0) ve alternatif hipotez (alternative hypothesis – H_1) aşağıda verilmiştir (Eşitlik 1).

$$H_0: \beta_U = \int_\theta B(\theta) f_F(\theta) d\theta = 0,$$
$$H_1: \beta_U = \int_\theta B(\theta) f_F(\theta) d\theta > 0 \quad (1)$$
$$B(\theta) = T_{SR}(\theta) - T_{SF}(\theta)$$

Burada β_U DMF miktarını belirleyen bir parametredir. Haliyle, boş hipotez DMF miktarının 0 olmasını, alternatif hipotez ise sıfırdan büyük olmasını belirtmektedir. $B(\theta)$ burada θ becerisi üzerinde

sırasıyla referans ve odak gruplarındaki öğrenciler arasındaki doğru cevap verme olasılıkları arasındaki farkı temsil ederken, $f_F(\theta)$ ise θ üzerine odak grubu için olasılık yoğunluk fonksiyonunu, $d\theta$ ise θ'nin diferansiyelini temsil etmektedir (Shealy ve Stout, 1993). Yani burada öğrenci grupları arasında θ becerisi anlamındaki fark integral yardımıyla taranmakta ve sonuç olarak her bir soru için DMF miktari belirlenmektedir. Belirlenen DMF miktarının sıfırdan anlamlı ölçüde büyük olma durumunda ise söz konusu madde DMF belirten madde olarak kodlanmaktadır.

Araştırmanın Etik İzinleri

Yapılan bu çalışmada "Yükseköğretim Kurumları Bilimsel Araştırma ve Yayın Etiği Yönergesi" kapsamında uyulması belirtilen tüm kurallara uyulmuştur. Yönergenin ikinci bölümü olan "Bilimsel Araştırma ve Yayın Etiğine Aykırı Eylemler" başlığı altında belirtilen eylemlerden hiçbiri gerçekleştirilmemiştir.

Etik kurul izin bilgileri:

Etik değerlendirmeyi yapan kurul adı = Kastamonu Üniversitesi Sosyal ve Beşeri Bilimler Araştırma ve Yayın Etik Kurulu,

Etik değerlendirme kararının tarihi = 25 Mart 2021,

Etik değerlendirme belgesi sayı numarası = E-16498365-050.01.04-2100025611.

Bulgular

Analiz Öncesi – Doğrulayıcı Faktör Analizi (DFA) (Confirmatory Factor Analysis – CFA)

DMF analizi gerçekleştirilmeden önce her bir kitapçıktaki soruların tek boyutluluğun incelenmesi gerekli olduğu yukarıda belirtilmişti (Jöreskog ve Sörbom, 1996). O yüzden, ilk olarak bu çalışmada kullanılan TIMSS 2015 ve TIMSS 2019 sınavlarının kitapçıkları (her bir sınav için 7 ayrı kitapçık) için Doğrulayıcı Faktör Analizi (DFA) gerçekleştirilmiştir. Tablo 5 bu analizlerin sonuçlarını göstermektedir.

Tablo 5. *TIMSS 2015 ve 2019 8. sınıf matematik sorularının kitapçıklar halinde Türk öğrenciler açısından DFA sonuçları*

Sınav	Grup	*N*	χ^2	df	CFI	TLI	RMSEA
TIMSS 2015	Kitapçık 1	435	917.060*	527	0.891	0.884	0.041
TIMSS 2015	Kitapçık 3	440	504.440*	350	0.943	0.939	0.032
TIMSS 2015	Kitapçık 5	440	819.410*	495	0.923	0.918	0.039
TIMSS 2015	Kitapçık 7	432	1,072.458*	464	0.834	0.823	0.055
TIMSS 2015	Kitapçık 9	435	505.506*	350	0.931	0.926	0.032
TIMSS 2015	Kitapçık 11	441	704.428*	377	0.889	0.880	0.044
TIMSS 2015	Kitapçık 13	435	652.800*	377	0.910	0.903	0.041
TIMSS 2019	Kitapçık 1	286	1,638.199*	740	0.765	0.753	0.065
TIMSS 2019	Kitapçık 3	287	1,780.549*	860	0.685	0.669	0.061
TIMSS 2019	Kitapçık 5	288	1,218.643*	464	0.747	0.729	0.075
TIMSS 2019	Kitapçık 7	287	1,723.573*	630	0.736	0.721	0.078
TIMSS 2019	Kitapçık 9	288	1,409.658*	594	0.739	0.723	0.069

TIMSS 2019	Kitapçık 11	289	1,185.374*	434	0.791	0.776	0.077
TIMSS 2019	Kitapçık 13	297	1,407.508*	741	0.788	0.776	0.058

*p<.001, CFI: Comparative Fit Index, TLI: Tucker-Lewis Index, RMSEA: Root Mean Square Error of Approximation

DFA analizinde tek boyutluluk adına bakılan ilk değer Ki-kare (χ^2) değeridir, ve bu değer için hesaplanan p-değerinin .05 değerinden büyük olması gerekir (Hooper, Coughlan ve Mullen, 2008). Ancak Ki-kare (χ^2) testinin varsayım olarak çok-değişkenli normalliği (multivariate normality) şart koşması ve çok değişkenli normallikten sapmaların esasında uygun olarak tanımlanmış bir modelin bile reddedilmesi gibi bir sonuç doğuracağını da belirtmişlerdir. Tablo 4'de görüldüğü üzere, iki sınavda da kullanılan her bir kitapçık için DFA modelleri için Ki-kare (χ^2) değerleri için hesaplanan p-değerleri örneklem büyüklüklerinin de etkisiyle anlamlı ölçüde .05 değerinden hatta .001 değerinden küçüktür (örn. TIMSS 2015 Kitapçık 1 için, $\chi^2(527) = 917.060$, $p<.001$). Bu durumda diğer model uyum indekslerine (goodness-of-fit) bakılmalıdır , örneğin CFI, TLI ve RMSEA (Hooper ve diğerleri, 2008). Hair, Black, Babin ve Anderson (2010) CFI gibi indekslerin özellikle büyük örneklemlerde .90 değerinden büyük olması gerektiğini belirtmektedir. Tablo 4'e bakıldığında, TIMSS 2015'de kullanılan Kitapçık 1, 7 ve 11 için CFI ve TLI değerlerinin bu kesme değerinden küçük olduğunu diğerleri için ise sorun olmadığını görebiliriz. TIMSS 2019 için ise bütün kitapçıklarda CFI değerlerinin bu kesme değerinden büyük olduğu görülmektedir. Burada CFI indeksi aslında TLI indeksinin revize edilmiş halidir (Hooper ve diğerleri, 2008). Hair ve diğerleri (2010) diğer bir yandan bu kesme değerinin çok karmaşık modeller için özellikle geçerli olduğunu belirtmişlerdir. Buradaki modellerde ise sadece soruların tek bir boyut altında olup olmadıkları söz konusudur. Bu yüzden, son olarak da RMSEA değerine göz atacak olursak, Hooper ve diğerleri (2008) bu değer için kesme değerini .08 olarak belirtmişlerdir. Aynı zamanda RMSEA değerinin modellemedeki tahmin edilen parametrelere olan hassasiyeti sebebiyle en çok bilgilendirici uyum endekslerinden biri olduğunu belirtmişlerdir. Haliyle, hem TIMSS 2015 sınavındaki bütün kitapçıklar (özellikle diğer indekslere göre tek boyutlu olamayacak Kitapçık 1, 7 ve 11) için hem de TIMSS 2019 sınavındaki tüm kitapçıklar için RMSEA değerleri bu değerden küçüktür. Sonuç olarak, TIMSS 2015 ve TIMSS 2019 sınavlarında kullanılan toplam 14 adet kitapçığın hepsinde kullanılmış olan soruların her bir kitapçık özelinde tek boyutluluklarının kabul edilebilir düzeyde olduğunu söyleyebiliriz.

Değişen Madde Fonksiyonu (DMF) Analizi Sonuçları

DMF analizi için gerekli olan tek boyutluluk şartı incelendikten sonra esas analiz olan DMF analizine geçebiliriz. Bir maddenin yanlılık belirtip belirtmediğini belirlemek için DMF prosedüründe kullanılan iki parametre, beta tahmini ve standartlaştırılmış p-fark indeksidir. Embretson ve Reise (2013), her iki grup için de madde zorluk parametresinin öncelikle her bir madde için hesaplandığı ve odak gruptaki her birinden ortalama fark çıkarılarak ayarlandığı beta tahmininin hesaplanma prosedürünü açıklamıştır. Daha sonra, her bir öğe için beta tahminine sahip olmak için referans ve ayarlanmış odak grubu öğe zorluk parametreleri arasındaki fark hesaplanır. Bu nedenle SIBTEST'teki

bu beta tahminleri, maddelerin referans grubu mu yoksa odak grubunu mu tercih ettiğini gösterir. Pozitif beta tahmini değerleri referans grubu (kız öğrenciler) lehine madde yanlılığı (DMF) olduğunu belirtirken, negatif değerler ise odak grubu (erkek öğrenciler) lehine yanlılık göstermektedir. Diğer bir yandan, standartlaştırılmış *p*-değerleri, her bir madde için referans ve odak grupları arasındaki toplam puan farkına bakılarak ve bu farklılıklar, toplam puanların her birindeki odak oranlarına göre ağırlıklandırılarak hesaplanır. Sonrasında bu *p*-değerleri, DMF işaretli öğe olup olmadığını belirtmek için .05 alfa seviyesi ile karşılaştırılır. Eğer elde edilen *p*-değeri .05'den küçükse, bu madde yanlılık (DMF) belirtiyor denilir ve boş hipotez (null hypothesis) reddedilir, yani referans ve odak gruplarının bu maddeye eşit şekilde doğru cevap verme olasılıkları yoktur. Ama, *p*-değeri .05'den büyükse, boş hipotez kabul edilir ve bu maddenin yanlılık belirtmediği sonucuna varılır (Shealy ve Stout, 1993). Tablo 6, TIMSS 2015 sınavında Tablo 7 ise TIMSS 2019 sınavında DMF belirten maddeleri ve bu maddeler için yukarıda belirtilen parametreleri göstermektedir.

Tablo 6. *Türkiye'de 8. sınıf düzeyinde cinsiyet üzerinden madde yanlılığı (DMF) belirten TIMSS 2015 matematik soruları.*

TIMSS Soruları	Yer aldığı kitapçık	Beta Tahmini	Standart Hata	χ^2	*p*-değeri
M042182	1	-0.1575	0.0643	5.9984	0.0143*
M042240	1	0.1226	0.0586	4.3721	0.0365*
M042164	1	0.1948	0.0674	8.3536	0.0038**
M062202	1	0.1753	0.0668	6.8925	0.0087**
M062115	3	-0.1172	0.0515	5.1740	0.0229*
M042023	5	0.2418	0.0893	7.3332	0.0068**
M042015	7	0.2041	0.0642	10.1057	0.0015**
M042114B	7	-0.1339	0.0619	4.6817	0.0305*
M042074A	7	0.1471	0.0701	4.4011	0.0359*
M042261	7	0.1764	0.0672	6.8916	0.0087**
M062244	7	0.2273	0.0741	9.4015	0.0022**
M062300	7	0.1605	0.0677	5.6195	0.0178*
M052413	9	0.1086	0.0539	4.0590	0.0439*
M052134	9	-0.1430	0.0586	5.9676	0.0146*
M062150	9	-0.1532	0.0538	8.1121	0.0044**
M062335	9	0.1377	0.0656	4.4022	0.0359*
M062133	9	0.1094	0.0557	3.8635	0.0493*
M052215	11	0.2160	0.0542	15.8642	0.0001***
M052067	11	0.1171	0.0596	3.8623	0.0494*
M062320	11	0.1565	0.0498	9.8833	0.0017**
M052125	13	0.1922	0.0529	13.1747	0.0003***
M052229	13	0.1552	0.0712	4.7576	0.0292*
M052063	13	0.1508	0.0557	7.3187	0.0068**
M052161	13	0.1489	0.0639	5.4381	0.0197*
M062192	13	0.1576	0.0489	10.3780	0.0013**

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

Tablo 6'da görüldüğü üzere TIMSS 2015 sınavında yer alan sekizinci sınıf matematik soruları üzerinde yapılan Değişen Madde Fonksiyonu (DMF) analizine göre, toplam 224 sorudan 25'inin yanlılık gösterdiği (DMF içerdiği) tespit edilmiştir. Bunlardan sadece 5 tanesi odak grubu (erkek

öğrenciler) lehine, kalan 20 tanesi ise referans grubu (kız öğrenciler) lehine yanlılık ifade etmektedir. Bu da demek oluyor ki, TIMSS 2015 sınavında yer alan 8. sınıf matematik sorularının sadece %2,2'si erkek öğrenciler lehine, %8,9'u ise kız öğrenciler lehine yanlılık belirtmektedir. Doğal olarak, toplamda tüm soruların %11,1'i madde yanlılığı (DMF) ifade etmektedir. Aşağıda benzer şekilde TIMSS 2019 sınavı üzerinde yapılan DMF analizinin sonuçları da verilmiştir.

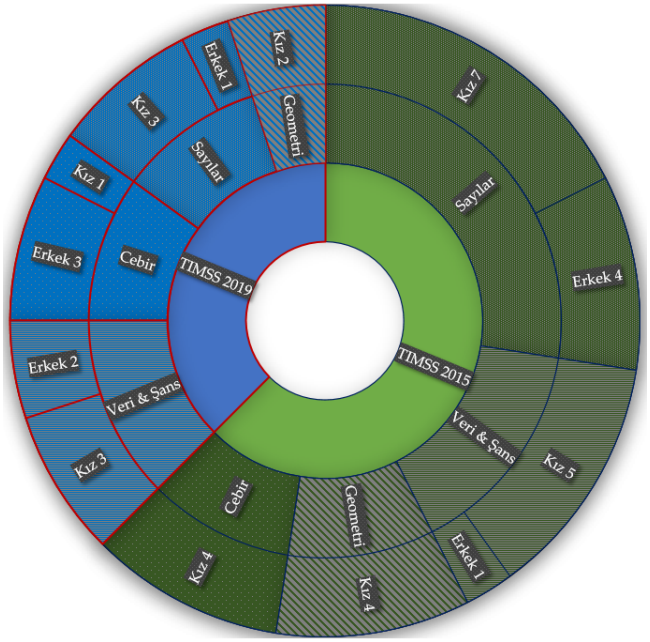
Tablo 7'de görüldüğü üzere TIMSS 2019 sınavında yer alan sekizinci sınıf matematik soruları üzerinde yapılan Değişen Madde Fonksiyonu (DMF) analizine göre, toplam 259 sorudan 15'inin yanlılık gösterdiği (DMF içerdiği) tespit edilmiştir. Bunlardan 6 tanesi odak grubu (erkek öğrenciler) lehine, kalan 9 tanesi ise referans grubu (kız öğrenciler) lehine yanlılık ifade etmektedir. Bu da demek oluyor ki, TIMSS 2019 sınavında yer alan 8. sınıf matematik sorularının %2,3'ü erkek öğrenciler lehine, %3,5'i ise kız öğrenciler lehine yanlılık belirtmektedir. Bu da, toplamda tüm soruların %5,8'inin madde yanlılığı (DMF) ifade ettiğini belirtmektedir.

Tablo 7. *Türkiye'de 8. sınıf düzeyinde cinsiyet üzerinden madde yanlılığı (DMF) belirten TIMSS 2019 matematik soruları.*

TIMSS 2019 Soruları	Yer aldığı kitapçık	Beta Tahmini	Standart Hata	χ^2 (ki-kare)	*p*-değeri
ME52125	1	0.3172	0.1577	4.0458	0.0443*
ME52229	1	0.4575	0.2318	3.8951	0.0484*
ME52146A	1	0.3286	0.1458	5.0783	0.0242*
ME72178C	3	0.3293	0.1212	7.3884	0.0066**
ME72027	3	-0.3409	0.1198	8.1005	0.0044**
ME52502B	5	0.2924	0.1249	5.4755	0.0193*
ME62345AB	9	0.2600	0.1250	4.3286	0.0375*
ME62171	11	0.2931	0.0988	8.8074	0.0030**
ME72221	11	-0.2315	0.0814	8.0892	0.0045**
ME62341	13	-0.2324	0.1068	4.7369	0.0295*
ME62242	13	-0.2468	0.1007	6.0088	0.0142*
ME72140D	13	0.2907	0.1287	5.1064	0.0238*
ME72154	13	-0.2860	0.1198	5.6974	0.0170*
ME72192	13	0.3018	0.1296	5.4232	0.0199*
ME72161	13	-0.3707	0.1415	6.8633	0.0088**

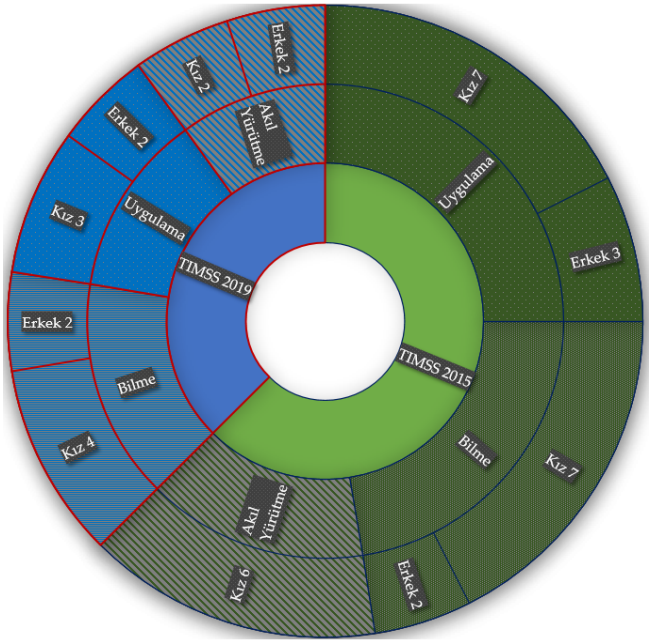
*$p<.05$, $p<.01$

Böylece, bu çalışmanın ilk araştırma sorusunun yanıtı ortaya çıkmıştır. İkinci araştırma sorusu için cevap bulmak üzere, aşağıda iki sınavda da madde yanlılığı tespit edilen soruların TIMSS kavramsal çerçevesinde tanımlanan içerik ve bilişsel alanlara göre dağılımları verilmiştir. Şekil 2 içerik alanlarına göre, Şekil 3 ise bilişsel alanlara göre TIMSS 2015 ve TIMSS 2019 sınavlarında yanlılık belirten soruların dağılımlarını göstermektedir. İki şekilde de dairesel bölgeler, belirtilen içerik veya bilişsel alanındaki soru sayısına orantılı olacak şekilde oluşturulmuştur. Bu sayede içerik alanlarına göre yanlılık belirten soru sayıları hem öğrencilerin cinsiyeti hem de sınav yılı anlamında eşzamanlı olarak karşılaştırılabilmektedir. Her bir bölgede bulunan soru sayıları da en dıştaki dairelerde belirtilmiştir.



Şekil 2. TIMSS 2015 ve 2019 sınavlarında Türkiye'de kız ve erkek öğrencilerin lehine yanlılık gösteren 8. sınıf matematik sorularının içerik alanlarına göre dağılımları.

Şekil 2'de görüldüğü üzere, TIMSS 2019 sınavında cinsiyete göre madde yanlılığı belirten soru sayısı TIMSS 2015 sınavındakine göre önemli ölçüde bir azalma göstermiştir. İçerik alanlarına ayrıntılı bakacak olursak, şekilde görüldüğü üzere 2015 yılında ne cebir ne de geometri içerik alanlarında erkek öğrenciler lehine yanlılık belirten soru bulunmamaktadır. Aynı yılda, sayılar ve veri ve şans alanlarında ise kız öğrenciler lehine yanlılık belirten soru sayısı erkek öğrencilere nispeten fazladır. 2019 yılına gelindiğinde ise kız ve erkek öğrenciler lehine yanlılık belirten soru sayılarında bir dengelenme görülmektedir. Buradaki dengenin kız öğrenciler lehine yanlılık belirten soru sayısındaki gözle görülür azalmadan kaynaklandığını söylemek mümkün olabilir. Yukarıda da belirtildiği üzere geometri içerik alanında yine erkek öğrenciler lehine yanlılık belirten soru bulunmamaktadır. Cebir içerik alanında 2015 yılında 4 soru kız öğrenciler lehine yanlılık belirtirken 2019 yılında sadece bir soru bulunmaktadır. Diğer bir yandan, 2015 yılında bu içerik alanında erkek öğrenciler lehine yanlılık belirten soru bulunmazken, 2019 yılında 3 soru erkek öğrenciler lehine yanlılık belirtmiştir. 2019 yılında erkek öğrenciler lehine sadece 6 sorunun yanlılık belirttiği düşünülürse, bu önemli bir değişimdir. Özetle, geometri içerik alanı iki sınavda da erkek öğrenciler lehine yanlılık belirtmemesi ile ön plana çıkarken; cebir içerik alanı ise 2015 yılında erkek öğrenciler lehine yanlılık belirtmezken, 2019 yılına gelindiğinde erkek öğrenciler lehine kız öğrencilere nispeten daha çok yanlılık belirten bir alan olarak gözlemlenmektedir. Şekil 3 ise belirtildiği üzere yine aynı soruların bu kez bilişsel alanlar yönünden dağılımını göstermektedir.



Şekil 3. TIMSS 2015 ve 2019 sınavlarında Türkiye'de kız ve erkek öğrencilerin lehine yanlılık gösteren 8. sınıf matematik sorularının bilişsel alanlara göre dağılımları.

Şekil 3'de görüldüğü üzere, TIMSS 2015 sınavında madde yanlılığı belirten soruların bilişsel alanlara göre dağılımı incelendiğinde, yine kız öğrenciler lehine yanlılık belirten soruların erkek öğrencilere nispeten bir ağırlığından söz etmek mümkündür. Özellikle akıl yürütme bilişsel alanında erkek öğrenciler lehine yanlılık belirten herhangi bir soru bulunmamaktadır. 2019 yılına gelindiğinde ise içerik alanlarına benzer şekilde, bilişsel alanlarda da yanlılık ifade eden soruların cinsiyetler arasında dengeli bir dağılım gösterdiği görülmektedir. Burada erkek öğrenciler lehine yanlılık belirten sorularda önemli ölçüde bir değişiklik bulunmamaktadır. 2019 yılına gelindiğinde oluşan bu dengeli durum, kız öğrenciler lehine yanlılık belirten soru sayısındaki önemli ölçüde gözlemlenen düşüştür. Bu şekillere ek olarak, Ek 1 ve Ek 2'de iki sınavda da yanlılık belirten soruların içerik ve bilişsel alanlara göre dağılımlarına ek olarak, içerik konuları ve soru başlıkları da verilmiştir. Bu konuda çalışmalar yapmak isteyecek araştırmacılar için önemli bir kaynak oluşturabilir.

Tartışma, Sonuç ve Öneriler

Matematik başarısı anlamında Türkiye genel olarak 2015 yılından 2019 yılına gelindiğinde 8. sınıf matematik alanında TIMSS sınavında 46 katılımcı arasındaki 29. sıradaki yerini (Mullis ve diğerleri, 2016) yine 46 katılımcı arasında 24. sıraya (Mullis ve diğerleri, 2020) yükseltmiştir. Başarı artışı gözlemlenirken kız ve erkek öğrenciler arasında genel ortalama puan anlamında iki sınavda da anlamlı ölçüde bir fark görülmemektedir (Tablo 2). Bu durum PISA 2006 sonuçları (Alacacı ve Erbaş, 2010; Demir ve diğerleri, 2010; Dinçer ve Kolasin, 2009) ve PISA 2009 sonuçları (Gürsakal, 2012) üzerine Türkiye hakkında yapılan çalışmaların bulgularıyla örtüşmemektedir. Bahsedilen çalışmalarda kız öğrencilerin erkek öğrencilere nispeten anlamlı ölçüde daha başarılı oldukları gözlemlenmiştir. Diğer bir yandan, TIMSS 2007 Türkiye sonuçları üzerine yapılan bir çalışmaya göre ise cinsiyet öğrenci başarısı açısından nötr bir gösterge olarak belirtilmiştir (Atar, 2011). Uluslararası veriler yerine

okullarda kendi örneklemlerini oluşturmak suretiyle yapılan çalışmalarda da yine benzer şekilde kız ve erkek öğrenciler arasında anlamlı bir başarı farkı gözlemlenmemiştir (Aksu, 2001; Işıksal ve Aşkar, 2005). Sonuç olarak, PISA Türkiye sonuçlarından farklı olarak TIMSS 2015 ve 2019 sınavlarında kız ve erkek öğrenciler arasında matematik performansı yönünden anlamlı bir farklılık yoktur. Ancak yine Tablo 2'de belirtildiği gibi, TIMSS 2015 ve 2019 sonuçlarına içerik ve bilişsel alanlar açısından bakacak olursak, hem katılımcı ülkeler genelinde hem de Türkiye özelinde kız ve erkek öğrenciler arasında anlamlı ölçüde farklar olduğunu gözlemleyebiliriz. Genel anlamda fark bulunmadığı halde içerik ve bilişsel alanlar yönünden ortaya çıkan bu farklılıklar, soruların yapısından kaynaklı yanlılık olma durumunu destekleyebilir.

Yürütülen bu çalışmanın bir sonucu olarak, TIMSS 2015 sınavında yer alan sekizinci sınıf matematik soruları üzerinde yapılan Değişen Madde Fonksiyonu (DMF) analizine göre, toplam 224 sorudan 25'inin yanlılık gösterdiği (DMF içerdiği) tespit edilmiştir. Bunlardan sadece 5 tanesi odak grubu (erkek öğrenciler) lehine, kalan 20 tanesi ise referans grubu (kız öğrenciler) lehine yanlılık ifade etmektedir. Ek olarak, TIMSS 2019 sınavında yer alan sekizinci sınıf matematik sorularında ise, toplam 259 sorudan 15'inin yanlılık gösterdiği (DMF içerdiği) tespit edilmiştir. Bunlardan 6 tanesi odak grubu (erkek öğrenciler) lehine, kalan 9 tanesi ise referans grubu (kız öğrenciler) lehine yanlılık ifade etmektedir. Bu çalışmada DMF analizi yapılırken Klasik Test Teorisine göre oluşturulan yöntemlerden ziyade Madde Tepki Kuramına göre oluşturulmuş olan SIBTEST prosedürü kullanılmıştır. Önceki araştırmalarda da belirtildiği üzere (Hambleton ve Jones, 1993; Kan ve diğerleri, 2013; Karakaya, 2012; Kelecioğlu ve diğerleri, 2014) bu şekilde daha hassas DMF ölçümleri yapıldığı da unutulmamalıdır. Sonuç olarak, TIMSS 2015 sınavındaki soruların %11,1'i (%2,2'si erkek, %8,9'u kız öğrenciler lehine), TIMSS 2019 sınavındaki soruların ise %5,8'i (%2,3'ü erkek, %3,5'i kız öğrenciler lehine) farklı cinsiyetlerde yanlılık belirtmiştir. Erkek öğrenciler açısından baktığımızda yanlılık belirten soru yüzdelerinde önemli bir değişiklik gözlemlenmezken, kız öğrenciler açısından durum önemli ölçüde bir azalma olarak karşımıza çıkmaktadır. Ancak 2019 yılında halen kız öğrenciler lehine yanlılık belirten soru sayısı erkek öğrencilerden fazladır. Böylece, içerik ve bilişsel alanlar yönünden bu soruların incelenmesi gerekliliği bir kez daha ortaya çıkmıştır.

Bu çalışmada yapılan DMF analizinin bir sonucu olarak TIMSS 2015 ve TIMSS 2019 sınavlarında kullanılan matematik sorularının Türkiye'deki 8. sınıf kız veya erkek öğrenciler lehine yanlılık belirtip belirtmediğine ek olarak yanlılık belirten soruların yine TIMSS kavramsal çerçevesinde belirtilen içerik ve bilişsel alanlar yönünden dağılımı da incelenmiştir (Şekil 2). İçerik alanları yönünden yapılan incelemeye göre,

– Geometri içerik alanı iki sınavda da erkek öğrenciler lehine yanlılık belirtmemesi ile ön plana çıkarken,

– Cebir içerik alanı ise 2015 yılında erkek öğrenciler lehine yanlılık belirtmezken, 2019 yılına gelindiğinde erkek öğrenciler lehine kız öğrencilere nispeten daha çok yanlılık belirten bir alan olarak gözlemlenmektedir.

Burada ilk olarak Bakan-Kalaycıoğlu ve Kelecioğlu'nun (2011) ÖSS sınavı üzerine yaptıkları DMF analizi çalışması akla gelmektedir. Çalışmalarında, 2005 yılında uygulanan ÖSS sınavında kullanılan 45 matematik sorusundan üç tanesinde yanlılık belirlemişlerdir. İki adet geometri sorusunun erkek öğrenciler lehine, bir adet cebir sorusunun ise kız öğrenciler lehine yanlılık belirttiğini bulmuşlardır. Ek olarak, Türkiye dışında bu alanda yapılan çalışmalarda da yine erkek öğrencilerin geometri, kız öğrencilerin de cebir alanında karşı cinslere göre daha başarılı oldukları belirtilmektedir (Lane ve diğerleri, 1996; McGraw ve diğerleri, 2006). TIMSS 2015 sınavında kullanılan 224 sorunun 43'ü, TIMSS 2019 sınavında ise 259 sorunun 54'ü geometri içerik alanına aittir (Şekil 1). İlginç bir şekilde bu iki sınavda da kullanılan hiçbir geometri sorusu erkek öğrenciler lehine yanlılık belirtmemektedir. Benzer şekilde, TIMSS 2015'de 64, TIMSS 2019'da da 69 soru cebir içerik alanına aittir. 2015 yılında da yine erkek öğrenciler lehine hiçbir cebir sorusu bulunmazken, 2019 yılında bu durum tersine dönmüş ve DMF belirten dört cebir sorusundan üçü erkek öğrenciler lehine yanlılık belirtmektedir. Bu anlamda, hem Bakan-Kalaycıoğlu ve Kelecioğlu'nun (2011) ulusal yerleştirme sınavı üzerindeki bulguları hem de Türkiye dışında yapılmış olan çalışmaların bulgularıyla (Lane ve diğerleri, 1996; McGraw ve diğerleri, 2006) bu çalışmanın TIMSS üzerine bulgularının örtüşmediği görülmektedir. Hem Türkiye hem de Türkiye dışındaki örneklemlerle yapılan çalışmalarla bu çalışmanın bulgularının örtüşmemesi TIMSS sınavında kullanılan soruların Türkiye'deki 8. sınıf kız ve erkek öğrenciler üzerinde farklı şekillerde etki edebileceğini ortaya koymaktadır.

Diğer bir yandan, TIMSS 2015 ve TIMSS 2019 sınavlarında kullanılan sorular aynı zamanda bilişsel alanlar açısından da incelenmiştir (Şekil 3). Sonuç olarak,

– TIMSS 2015 sınavında, kız öğrenciler lehine yanlılık belirten soruların erkek öğrencilere nispeten bir ağırlığından söz etmek mümkündür. Özellikle akıl yürütme bilişsel alanında erkek öğrenciler lehine yanlılık belirten herhangi bir soru bulunmamaktadır.

– 2019 yılına gelindiğinde ise içerik alanlarına benzer şekilde, bilişsel alanlarda da yanlılık ifade eden soruların cinsiyetler arasında dengeli bir dağılım gösterdiği görülmektedir.

Burada akıl yürütme alanı erkek öğrenciler lehine 2015 yılında hiçbir soru içermemesi ile ön plana çıkmıştır. Bu durum içerik alanlardaki durumun aksine önceki çalışmaların bulgularıyla örtüşmektedir. Zira kız öğrencilerin akıl yürütme problemlerinde erkek öğrencilere nispeten daha başarılı oldukları önceki araştırmacılar tarafından ortaya konulmuştur (Friedman, 1996; Ryan ve Chiu, 1996). Ek olarak, 2015 yılında ortalama puan olarak üç bilişsel alanda da Türkiye'de kız ve erkek öğrenciler arasında TIMSS sınavında anlamlı bir puan farkı gözlemlenmemiştir (Tablo 2). Ancak 2019 yılına gelindiğinde bilme ve akıl yürütme bilişsel alanlarında istatistiksel olarak kız öğrencilerin anlamlı

ölçüde erkek öğrencilere nispeten daha yüksek ortalama puana sahip oldukları görülmektedir. Diğer bir yandan ise TIMSS 2015'de bilişsel alanlar yönünden kız öğrenciler lehine daha çok yanlılık belirten soru bulunmaktayken, 2019'da bu durum dengelenmiştir.

İster TIMSS 2015 isterse de TIMSS 2019 sınavları olsun, bu sınavlarda kullanılan soruların içerikleri maalesef tamamıyla açık veri olarak paylaşılmamaktadır. Bu açıdan, özellikle bu çalışmanın bulguları doğrultusunda geometri ve cebir içerik alanları ile akıl yürütme bilişsel alanı Türkiye'de kız ve erkek öğrenciler arasında yanlılık belirtme durumu açısından daha detaylı şekilde ele alınmalıdır. Bu alanların özellikle ders kitaplarında kullanılan sorular itibariyle cinsiyet açısından inceleneceği ve bu alanlara yönelik kız ve erkek öğrencilerin tutumlarına yönelik yapılacak araştırmalar alan yazına katkı sağlayacaktır. Özellikle de Türkiye hem de Türkiye dışında örneklemler üzerinde yapılan çalışmalar neticesinde geometri içerik alanı erkek öğrencilerin cebir içerik alanı ise kız öğrencilerin daha başarılı oldukları veya yanlılık belirten soruların bu minvalde yanlılığa sahip oldukları alanlar olarak gözlemlenmekte iken TIMSS 2015 ve TIMSS 2019 sonuçları bu durumun tam olarak aksini ortaya koymuştur. Bu anlamda özellikle de bu iki içerik alanının ders kitaplarında, derslerin işlenişi sırasında, vs. farklı cinsiyetler açısından nasıl ele alındığının ayrıntılı inceleneceği çalışmalar bu çalışmanın devamında faydalı olacaktır. Ek 1 ve Ek 2'de bu çalışmada yanlılık belirttiği sonucuna varılan soruların içerik ve bilişsel alanlar yönünden dağılımına ek olarak, yine bu soruların içerik konuları ve soru başlıkları da sağlanmıştır. Bu konuda yapılacak olan çalışmalara katkı sağlaması bakımından dileyen araştırmacılar çekinmeden kullanabilirler.

Ahi Evran University
Journal of Kırşehir Education Faculty

ISSN: 2147 - 1037

http://kefad.ahievran.edu.tr

ENGLISH VERSION

Introduction

International assessments such as the Trends in International Mathematics and Science Study (TIMSS) created by the International Association for the Evaluation of Educational Achievement (IEA) and the Programme for International Student Assessment (PISA) created by the Organization for Economic Co-operation and Development (OECD) show countries their performance in mathematics and science among other participating countries (Akyüz, 2014; Akyüz and Berberoğlu, 2010; Doğan and Barış, 2010; İncikabı, 2012). These assessments have been used since the 1960s (Yıldırım, Yıldırım, Ceylan, Yetişir and Ajans, 2013). The results obtained from international educational assessments provide countries a perspective in order to shape their education policies to improve the performance of students depending on how the results are interpreted (Akyüz, 2006; Akyüz, 2014; Akyüz and Berberoğlu, 2010; Bilican, Demirtaşlı and Kilmen, 2011; Doğan and Barış, 2010; İncikabı, 2012). For this reason, countries participate in these international assessments to monitor the progress of their students' achievement in the international environment and to investigate the factors affecting their achievement (Akyüz, 2014; Doğan and Barış, 2010; İncikabı, 2012). By understanding these factors and comparing education systems with others, education policymakers can evaluate their decisions, identify problems and develop more effective policies (Akyüz and Berberoğlu, 2010; Bilican et al., 2011).

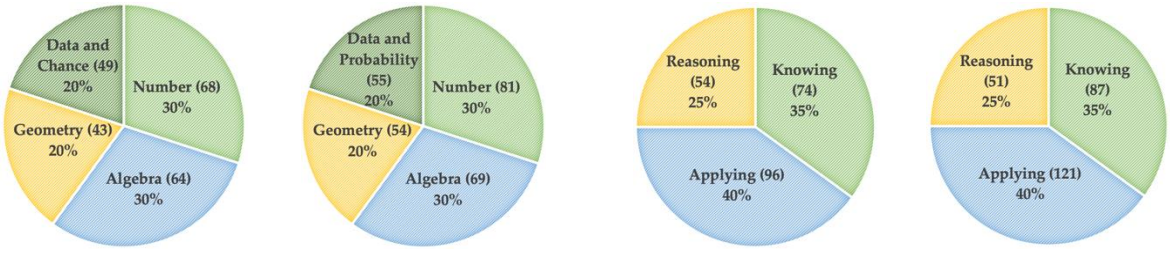
TIMSS is an international assessment repeated every four years to evaluate the performance of 4th and 8th grade students in participating countries around the world in the fields of mathematics and science (Doğan and Barış, 2010; İncikabı, 2012). Turkey participated in this assessment for the first time in 1999. While Turkey participated in the evaluation in 1999 only at the eighth-grade level, Turkey did not participate in 2003 and participated in 2007 at the eighth-grade level again. By 2011, Turkey has started to participate in both the fourth- and eighth-grade levels (Bilican et al., 2011; Erkan, 2013; Güner, Sezer and İspir, 2013) and has continued to participate in 2015 (Mullis, Martin, Foy and Hooper, 2016) and 2019 (Mullis, Martin, Foy, Kelly and Fishbein, 2020). Unlike the other years, in 2019, Turkey also participated in this assessment electronically. Turkey's position relative to other participants during the years of participation is given in Table 1.

Table 1. *Turkey's position in TIMSS assessments among other participants across years*

	4th grade		8th grade	
	Ranking of Turkey	Number of Participants	Ranking of Turkey	Number of Participants
TIMSS 1999 (Akyüz, 2006)	-	-	31	38
TIMSS 2007 (Martin, Mullis and Foy, 2008)	-	-	37	67 (8)
TIMSS 2011 (Mullis, Martin, Foy and Arora, 2012)	40	60 (8)	37	59 (14)
TIMSS 2015 (Mullis et al., 2016)	41	57 (7)	29	46 (7)
TIMSS 2019 (Mullis et al., 2020)	26	64 (6)	24	46 (7)

Note: The numbers given in parentheses indicate the regions that specifically participated from some countries among the total participants (*benchmarking participants*).

As can be seen in Table 1, Turkey has relatively improved its position over the years. Especially with its performance in 2015 and 2019, Turkey has entered a development trend compared to other years. It is especially important to be able to meaningfully read the changes in these two years in order to maintain the increase in the following years. In this sense, in the mathematics section of the TIMSS international assessment, which is organized to reveal the achievement of the students among other countries and the reasons behind the success or failure, the development of student success in both content and cognitive domains can be observed. In this assessment, the items are formed in terms of certain content and cognitive domains. While the content domains of the eighth-grade items were *numbers, algebra, geometry,* and *data and chance* in the TIMSS 2015, the data and chance sub-content domain was changed to *data and probability* in the TIMSS 2019. On the other hand, cognitive domains are stated as *knowing*, *applying,* and *reasoning in* both years (Gronmo, Linquist, Arora and Mullis, 2013; Martin, Mullis and Foy, 2017). The targeted percentage distributions of these content and cognitive domains in terms of the items and the actual number of items (in parentheses) in the TIMSS 2015 and TIMSS 2019 are given below (Figure 1). There are 224 eighth-grade mathematics items in TIMSS 2015 and 259 eighth-grade mathematics items in TIMSS 2019. The items in TIMSS 2019 are the ones used in the newly created electronic assessment.



Şekil 1. The targeted (%) and actual (in parentheses) distributions of the 8th grade mathematics items used in TIMSS 2015 and TIMSS 2019 in terms of content and cognitive domains (derived from Gronmo et al., 2013 and Martin et al., 2017)

Gender and Mathematics Achievement

The factors affecting student achievement is one of the subjects that researchers are interested in. In particular, factors related to students and families (e.g., Akyüz, 2014; Atar, 2011; Demir, Kılıç and Ünal, 2010), factors related to school (e.g., Demir et al., 2010; Engin-Demir, 2009; Kılıç, Çene and Demir) , 2012), and the ones related to teachers (e.g., Aaronson, Barrow and Sander, 2007; Clotfelter, Ladd and Vigdor, 2007; Clotfelter, Ladd and Vigdor, 2010; Hill, Rowan and Ball, 2005; Nye, Konstantopoulos and Hedges, 2004; Stronge, Ward and Grant, 2011) have attracted the attention of researchers. Among these factors, the student gender has also taken its place in the literature as an important factor affecting achievement. Especially in studies carried out to model student achievement, the gender factor was considered as a control variable and was included in the analyses by controlling its effect on student achievement while examining the effect of other factors in terms of student achievement (e.g., Aaronson et al., 2007; Boyd, Grossman, Lankford, Loeb, and Wyckoff). , 2005; Clotfelter et al., 2010; Goldhaber and Anthony, 2007; Hill et al., 2005; Jacob and Lefgren, 2002; Rowan, Chiang and Miller, 1997; Stronge et al., 2011). In studies conducted in Turkey, on the other hand, gender creates a dilemma in terms of mathematics achievement. While some researchers have determined that student gender is a factor that plays an important role in mathematics achievement (Alacacı and Erbaş, 2010; Demir et al., 2010; Dinçer and Kolasin, 2009; Gürsakal, 2012; Kılıç et al., 2012), some researchers stated opposite results indicating that there is no connection between gender and mathematics achievement (Aksu, 2001; Atar, 2011; Işıksal and Aşkar, 2005). Studies generally considered gender as a factor affecting student achievement and tried to explain student success in terms of gender along with other variables.

In order to see the differences in mathematics performance between the genders, the average achievement scores of all participant countries by gender are given below in terms of both content and cognitive domains, based on TIMSS 2015 and TIMSS 2019 data (Table 2). Table 2 also gives the mean scores for Turkey in these years.

Table 2. *Turkey and TIMSS overall mean achievement scores in content and cognitive domains in terms of gender at eighth-grade in TIMSS 2015 and TIMSS 2019*

		Turkey 2015		TIMSS 2015		Turkey 2019		TIMSS 2019	
		Female	Male	Female	Male	Female	Male	Female	Male
Content Domains	Numbers	461	455	484	491*	496	490	493	497*
	Algebra	443	452*	492*	481	503*	482	503*	493
	Geometry	469*	450	487*	480	496*	483	499*	495
	Data and Chance	472*	454	481*	479	506	498	490	489
Cognitive Domains	Knowing	470	464	488*	485	503*	485	499*	494
	Applying	450	444	486	486	494	488	497	496
	Reasoning	461	458	487*	482	511*	497	501*	497
	Overall Mean	461	455	488*	485	501	490	491	488

*Significantly higher than the other gender (at .05 level)

(Derived from the National Center for Education Statistics, 2021 and Mullis et al., 2020)

As seen in Table 2, when looking at the overall mean scores, no statistically significant difference was observed between the mean scores of male and female students neither in terms of all TIMSS

participants nor in terms of Turkey (except for TIMSS 2015). In terms of content and cognitive domains, it has been observed that female students generally have relatively higher mean scores in content and cognitive domains than male students, in terms of the mean score of all countries participated in the TIMSS. Looking at the mean scores of all participating countries, both in 2015 and 2019, male students only outperformed female students significantly on average in the number content domain in 2015. In terms of cognitive domains, it was observed that female students were statistically significantly more successful than male students in cognitive domains of knowing and reasoning, while male students were relatively more successful than female students in the field of applying sub-domain.

When Table 2 is analyzed for Turkey in terms of content domains, female students have significantly higher mean scores than male students in geometry and data and chance content domains while male students have higher mean scores than female students in the algebra sub-content domain. In 2015, no significant difference was observed between male and female students in Turkey in the cognitive domains. When it comes to 2019, it is seen that female students have higher mean scores in content and cognitive domains than male students in Turkey, similar to the other participant countries. However, male students did not have significantly higher scores than female students in any of the content and cognitive domains. In addition, there is no significant difference between these two genders in terms of overall mean scores of either in Turkey or other countries, but there is a positive trend especially for female students in the content and cognitive domains. In this sense, it is important to read the achievement differences between the genders in terms of content and cognitive domains.

When looking at the previous studies in this sense, according to a study conducted on 69 countries that participated in the TIMSS and PISA, it was seen that male students exhibited relatively more positive attitudes towards mathematics than female students although there was not much difference indicated between them in terms of achievement (Else-Quest, Hyde and Linn, 2010). The important question is whether this attitude was reflected in their achievement? In this respect, it is necessary to mention the other studies carried out in the content and cognitive domains of mathematics. In terms of content domains, studies were stating that male students were more successful than female students in geometry, and female students were more successful in algebra than male students (Lane, Wang and Magone, 1996; McGraw, Lubienski and Strutchens, 2006). In terms of cognitive domains, previous researchers have shown that female students were more successful than male students in reasoning problems (Friedman, 1996; Ryan and Chiu, 1996).

It is also necessary to mention the studies conducted using the Turkish sample. When looking at the studies examining the relationship between gender and mathematics achievement, especially those who use the data of international exams, Alacacı and Erbaş (2010) using Hierarchical Linear Modeling (HLM) on PISA 2006 results found that gender was an important indicator of success. The results they obtained showed that male students were more successful than female students. Demir et

al.'s (2010) study on PISA 2006 results, again using HLM, supported this finding that male students had better scores in mathematics. Dinçer and Kolasin (2009), in addition, found that female students had an average score of 14 points lower than male students in the mathematics test in PISA 2006; however, they stated that they had an average of 32 points higher than male students in the reading test. In PISA 2009, the results regarding the gender of students were again similar for Turkey. The gender of the students has a significant effect on the mathematics achievement of the students (Kılıç et al., 2012). In addition to their success, it was stated that male students were expected to perform higher than female students in terms of self-expectations in the PISA 2009 (Gürsakal, 2012).

On the other hand, according to the model created based on Item Response Theory (IRT) in TIMSS 2007, the gender variable was found to be a neutral indicator of students' mathematics achievement in Turkey (Atar, 2011). Again, according to Aksu's (2001) study on students studying at a private school in Ankara, no significant relationship was found between the gender of the students and their performance. In addition, Işıksal and Aşkar's (2005) study conducted on 64 seventh-grade students in Turkey stated that there was no significant difference between the genders in terms of average scores neither in terms of their mathematics performance nor of their mathematical self-confidence.

In summary, results of some studies indicated that male students were more successful especially in the PISA in Turkey than female students (Alaçacı and Erbaş, 2010; Demir et al., 2010; Dinçer and Kolasın, 2009; Kılıç et al., 2012; Gürsakal, 2012) while some other studies stated that there was no significant connection between student gender and their achievement (Aksu, 2001; Atar, 2011; Işıksal and Aşkar, 2005). On the other hand, when the results of TIMSS 2015 and TIMSS 2019 were examined in terms of overall mean scores, no significant difference was observed between female and male students. However, it was seen that female students had relatively significantly higher scores than male students in terms of the mean scores calculated for content and cognitive domains across the participant countries or in Turkey in particular. According to previous studies, it has been observed that male students are more successful in geometry and female students in algebra content (Lane et al., 1996; McGraw et al., 2006). In addition, it was stated by previous researchers that female students are more successful than male students in the reasoning cognitive domain (Friedman, 1996; Ryan and Chiu, 1996).

Gender Bias in Exams

One of the ways to understand the difference between the mathematics achievement of female and male students is to examine the structure of the items used in the exams and to identify possible biases. As mentioned above, the mathematics performances of these two groups of students, measured with different measurement tools, differ. The possibility that the performance differences between these two genders stem from the structure of the items used in the assessments should not be ignored. According to the Differential Item Functioning (DIF) analysis conducted by Bakan-Kalaycıoğlu and Kelecioğlu (2011) to determine the items showing bias between genders on the Student Selection

Examination (ÖSS) administered in Turkey in 2005, three mathematics items showed potential bias in the mathematics subtest consisting of 45 items, while one of them was an algebra item, two of them were geometry. Detection of DIF indicates the possibility that the items may show bias towards one of the groups. As a result, an algebra item showed DIF in favor of female students, while two geometry items showed DIF in favor of male students. DIF analysis is a method used in this study as well, and it is a special statistical method used to detect items that may indicate bias. In the method section of the study, detailed information about this analysis technique is provided.

In addition, Karakaya (2012) determined DIF in two items according to the DIF analysis they conducted on the mathematics items in the Placement Exam (SBS), which was applied in 2009 and which eighth-grade students attend in order to enter high schools in Turkey. However, as a result of expert opinion, these items were not in favor of any gender in terms of item bias. In the mentioned study, the Mantel-Haenszel method, which was created according to the Classical Test Theory (CTT), was preferred. Again, in another DIF study applied for the SBS exam in 2009, three different DIF analysis methods were used to examine whether the mathematics items indicate bias in terms of gender and school types (Kelecioğlu, Karabay and Karabay, 2014). In the aforementioned study, it was concluded that DIF was stated in 14 out of 20 mathematics items, but it was concluded that 3 of them did not constitute a bias as a result of expert opinion. From this point of view, the importance of the method used in the same analysis on the items of the same exam emerges. In the analysis of Kan, Sünbül and Ömür (2013) based on Item Response Theory (IRT) rather than CTT, again on the SBS exam but with the format applied in 2011, 20 of the 20 math items according to the Lord's method were also analyzed. According to the Raju method, they found a bias in favor of male or female students in 15 of them. While CTT is based on a reporting system on students' scores or predicted scores in a test, on the contrary, IRT is based on the logic of making connections over students' skill levels and the characteristic values of each item in the test (Hambleton and Jones, 1993). Hambleton and Jones (1993) also stated that for these two theories, the assumptions required for CTT were weak, that is, it is easier to provide the assumptions necessary to make analyzes for this theory, but the assumptions required for IRT were much stronger. For this reason, an IRT-based DIF analysis seems to be more sensitive when detecting possible biases in items.

As a result, while no significant difference was observed between genders in terms of mathematics performance in experimental studies conducted by creating a sample in Turkey, differences were revealed between these two genders in terms of mathematics performance, especially in national and international exams (especially TIMSS and PISA). In this sense, the observation of these differences, which are not observed in experimental studies, in studies at the national or international level may also be related to the structure of the items used in the exams. On the other hand, according to TIMSS and PISA results in Turkey, the success of male or female students varies. In this sense, according to the studies, the results of which are given in detail above, and which examine the biases of

different genders in terms of item structures, it was seen that mathematics items showed bias in exams used for placement purposes, especially in Turkey. As a matter of fact, in the studies indicating bias on the national exams, the performance differences between the genders in Turkey that may be caused by the item structures have been examined by the researchers. However, there were not enough studies that reveal these biases by using international exam data on students in Turkey. Considering the differences in achievement among students in different genders in these areas (Friedman, 1996; Lane et al., 1996; McGraw et al., 2006; Ryan et al. Chiu, 1996), the need for a study that reveals the biases of the items used in the assessments in terms of the content and cognitive domains specified in the assessments emerges more clearly.

For this purpose, in this study, it was aimed to determine the item biases of the eighth-grade mathematics items used in the TIMSS 2015 and TIMSS 2019 international assessments in Turkey in terms of male and female students. In other words, it was aimed to reveal the situations where the items used in the assessments yield advantages or disadvantages for students of different genders. Although it is accepted that these two groups of students theoretically have an equal level of knowledge on the structures that the assessment items aim to measure, differences can be observed in the answers due to the structure of the items. Especially the differences in the approaches of male and female students to the mathematics lesson and the closeness of the expressions used in the items to these different approaches may play a role in the different behavior of the items between these two genders. Therefore, this study aimed to reveal the eighth-grade mathematics items of TIMSS 2015 and TIMSS 2019, which work differently on groups of female and male students in Turkey, with a special analysis method, Differential Item Functioning (DIF). As a result of this analysis, it was aimed to determine the mathematics items that functioned biasedly for male and female students and to reveal whether these items have a distinguishable feature in terms of the content and cognitive domains determined in the TIMSS assessment. Although it is not possible to say that the biases in the items may only be caused by gender, it would be helpful to examine the content and cognitive domains of these items in more detail in terms of gender. Thus, the study is shaped around the following research questions:

– Do eighth-grade mathematics items used in TIMSS 2015 and TIMSS 2019 international assessments create a bias between male and female students in Turkey?

– If there are items found to be biased, what are the characteristics of these items in terms of content and cognitive domain specified in the TIMSS 2015 and TIMSS 2019 conceptual frameworks?

Method

Data Collection

The sample of the study consisted of the participants in the mathematics section of the eighth-grade TIMSS 2015 (n_{female} = 1,577, n_{male} = 1,481, n_{total} = 3,058) and TIMSS 2019 (n_{female} = 1,027, n_{male} = 995, n_{total} = 2,022) assessments in Turkey, total of 5,080 students who answered the odd-numbered item

booklets. Based on the answers given by these students to a total of 438 different mathematics items, 224 for TIMSS 2015 and 259 for TIMSS 2019, it was tried to identify biased items for female or male student groups.

Students who took the TIMSS 2015 and TIMSS 2019 assessments answered one of the 14 different booklets. There are two of the 14 different math item blocks in each booklet. The math item blocks that students answered in both assessments, according to the booklets they received, and the number of items in these blocks are shown below (Table 3).

Table 3. *TIMSS 2015 and TIMSS 2019 eighth-grade booklets in terms of math blocks and number of items*

	Item Blocks		**TIMSS 2015**			**TIMSS 2019**		
	Block-1	Block-2	Block-1	Block-2	Total	Block-1	Block-2	Total
Booklet 1	M01	M02	17	18	35	16	24	40
Booklet 2	M02	M03	18	15	33	24	16	40
Booklet 3	M03	M04	15	18	33	16	27	43
Booklet 4	M04	M05	18	19	37	27	18	45
Booklet 5	M05	M06	19	15	34	18	14	32
Booklet 6	M06	M07	15	17	32	14	16	30
Booklet 7	M07	M08	17	15	32	16	21	37
Booklet 8	M08	M09	15	15	30	21	18	39
Booklet 9	M09	M10	15	14	29	18	18	36
Booklet 10	M10	M11	14	15	29	18	15	33
Booklet 11	M11	M12	15	14	29	15	16	31
Booklet 12	M12	M13	14	16	30	16	16	32
Booklet 13	M13	M14	16	16	32	16	25	41
Booklet 14	M14	M01	16	17	33	25	16	41

(Derived from Martin, Mullis and Foy, 2013 and Martin et al., 2017)

As shown in Table 3, it is possible to examine all possible math item blocks when either odd- or even-numbered booklets are examined. Starting from M01 to M14, there are a total of 14 different math item blocks. If odd- or even-numbered booklets are used for both TIMSS 2015 and TIMSS 2019, it would be possible to examine all possible items in terms of item bias. Therefore, it is preferred to use odd-numbered booklets, although there is no difference. Table 4 shows the number of students who answered each odd-numbered booklet and their gender distribution.

Table 4. *TIMSS 2015 and TIMSS 2019 mathematics booklets and the gender distribution of eighth-grade students who answered in Turkey (sample of the study)*

Booklet and Item Blocks		Number of Items		Number of Students					
		TIMSS 2015	TIMSS 2019	TIMSS 2015			TIMSS 2019		
				Female	Male	Total	Female	Male	Total
Booklet 1	M01-M02	35	40	228	207	435	157	129	286
Booklet 3	M03-M04	33	43	229	211	440	144	143	287
Booklet 5	M05-M06	34	32	229	211	440	139	149	288
Booklet 7	M07-M08	32	37	221	211	432	152	135	287
Booklet 9	M09-M10	29	36	218	217	435	140	148	288
Booklet 11	M11-M12	29	32	218	223	441	147	142	289
Booklet 13	M13-M14	32	39	234	201	435	148	149	297
	Total	224	259	1,577	1,481	3,058	1,027	995	2,022

Data Analysis

The study is quantitative in nature, and the Differential Item Functioning (DIF) technique was used to determine the gender bias of the items. Before performing DIF, the math items in each group should be in a unidimensional structure, as an assumption (Jöreskog and Sörbom, 1996). For this reason, firstly, Confirmatory Factor Analysis (CFA) was performed using the *lavaan* package (Rosseel, 2012) over the R Studio open-source statistical program (R Core Team, 2018). After examining the uni-dimensionality of the items in each booklet, it was used *difSIBTEST* module (Shealy and Stout, 1993) in *difR* package (Magis, Beland, Tuerlinckx and De Boeck, 2010) to conduct DIF analysis in R environment.

Item bias is the probability that students in different groups may show different performances on the items even if they theoretically have equal skills on the latent variable that the items want to measure (Zumbo, 1999). In other words, although male and female students in Turkey have equal expectations of success in assessments, it may be possible due to the nature of the problem that some items are biased towards one of these genders. Thanks to the DIF analysis technique, it is possible to identify the mathematics items that are biased towards either male or female students. It is also important to examine the identified items in terms of their content and cognitive domains.

In the world of Item Response Theory (IRT), Differential Item Functioning (DIF) analysis is a statistical method that replaces the concepts of the item or test deviation (see Embretson and Reise, 2013; Zumbo, 1999). DIF occurs when a candidate item differs between groups in terms of having the same relationship with the latent variable that the item wants to measure (see Embretson and Reise, 2013). In addition, Embretson and Reise (2013) stated the multidimensional SIBTEST (Simultaneous Item Bias Test) model created by Shealy and Stout (1993) for DIF as an important development in DIF analysis. Shealy and Stout (1993) defined the groups compared in the DIF analysis as reference and focal groups. Compared to other DIF determination methods, the SIBTEST procedure emerges as a newer method and is based on the system of arranging the mean score values obtained by the Mantel-Haenszel method, which is widely used, by linear regression in line with the total score that individuals generally obtain from the test (Osterlind and Everson, 2009). Shealy and Stout (1993) stated that the method they created, the procedure of SIBTEST, can examine DIF in a test with several items and provides an opportunity to cognitively examine bias/DIF in items. In addition, Awuor (2008) states that SIBTEST is a non-parametric procedure developed as an extension of Shealy and Stout's (1993) study and uses actual item response data instead of estimated parameters. Shealy and Stout (1993) stated that SIBTEST can detect DIF throughout a test with more than one item or for only one item. Embretson and Reise (2013) found the SIBTEST procedure very pleasing to the ear and stated that the theory behind the model fitted real-world tests very well and would attract more attention from researchers in the near future. As an extension of the work of Shealy and Stout (1993), the computer program SIBTEST was developed by Stout and Roussos (1996). This computer program later took its place as a module named difSIBTEST

in the difR package, which is a comprehensive package developed for DIF analysis in the R Studio statistical program. This module was also used for this study.

In this study, while the reference group was female students, the focal group was male students who participated in the TIMSS 2015 and TIMSS 2019 international assessments at the eighth-grade level in Turkey. Embretson and Reise (2013) stated that these groups (reference and focal) may have different mean and standard deviation values over the latent variable. However, they stated that these differences were not signs of DIF and might complicate the DIF analysis procedure depending on the research setting. However, Zumbo (1999) pointed out that these two groups should be matched on the relevant trait variable in order to apply the DIF analysis in terms of determining the probability of answering the item correctly. Another important consideration when applying DIF analysis on a latent trait variable is that researchers interested in identifying DIF use the same items for reference and focal groups (see Embretson and Reise, 2013). The null hypothesis (H_0) and alternative hypothesis (H_1) accepted for the DIF analysis are given below (Equation 1).

$$H_0: \beta_U = \int_\theta B(\theta) f_F(\theta) d\theta = 0,$$
$$H_1: \beta_U = \int_\theta B(\theta) f_F(\theta) d\theta > 0 \quad (1)$$
$$B(\theta) = T_{SR}(\theta) - T_{SF}(\theta)$$

In Equation 1, β_U parameter determines the amount of DIF. Thus, the null hypothesis states that the amount of DIF should be 0 while the alternative hypothesis states it should be greater than 0. $B(\theta)$ represents the difference between the probability of giving correct answers between the students in the reference and focal groups in terms of the θ skill. In addition, $f_F(\theta)$ represents the probability density function for the focal group and $d\theta$ represents the differential of θ (Shealy and Stout, 1993). In other words, here, the difference between the student groups in terms of the skill θ is scanned with the help of integral and as a result, the amount of DIF for each item is determined. If the determined DIF amount is significantly greater than zero, the item is coded as indicating DIF.

Ethical Permissions of Research

In this study, all the rules specified to be followed within the scope of the "Higher Education Institutions Scientific Research and Publication Ethics Directive" were complied with. None of the actions specified under the title of "Actions Contrary to Scientific Research and Publication Ethics", which is the second part of the directive, were carried out.

Ethics committee permission information:

Name of the committee that made the ethical evaluation = Kastamonu University Social and Human Sciences Research and Publication Ethics Committee,

Date of ethical review decision = 25 March 2021,

Ethical document issue number = E-16498365-050.01.04-2100025611.

Results

Pre-Analysis – Confirmatory Factor Analysis

It was stated above that it was necessary to examine the unidimensionality of the items in each booklet before performing the DIF analysis (Jöreskog and Sörbom, 1996). Therefore, firstly, Confirmatory Factor Analysis (CFA) was performed for the booklets of the TIMSS 2015 and TIMSS 2019 assessments (7 separate booklets for each exam) used in this study. Table 5 shows the results of these analyzes.

Table 5. *CFA results of TIMSS 2015 and 2019 eighth-grade mathematics items for Turkish students*

Assessment	Booklet	*N*	χ^2	df	CFI	TLI	RMSEA
TIMSS 2015	Booklet 1	435	917.060*	527	0.891	0.884	0.041
TIMSS 2015	Booklet 3	440	504.440*	350	0.943	0.939	0.032
TIMSS 2015	Booklet 5	440	819.410*	495	0.923	0.918	0.039
TIMSS 2015	Booklet 7	432	1,072.458*	464	0.834	0.823	0.055
TIMSS 2015	Booklet 9	435	505.506*	350	0.931	0.926	0.032
TIMSS 2015	Booklet 11	441	704.428*	377	0.889	0.880	0.044
TIMSS 2015	Booklet 13	435	652.800*	377	0.910	0.903	0.041
TIMSS 2019	Booklet 1	286	1,638.199*	740	0.765	0.753	0.065
TIMSS 2019	Booklet 3	287	1,780.549*	860	0.685	0.669	0.061
TIMSS 2019	Booklet 5	288	1,218.643*	464	0.747	0.729	0.075
TIMSS 2019	Booklet 7	287	1,723.573*	630	0.736	0.721	0.078
TIMSS 2019	Booklet 9	288	1,409.658*	594	0.739	0.723	0.069
TIMSS 2019	Booklet 11	289	1,185.374*	434	0.791	0.776	0.077
TIMSS 2019	Booklet 13	297	1,407.508*	741	0.788	0.776	0.058

*p<.001, CFI: Comparative Fit Index, TLI: Tucker-Lewis Index, RMSEA: Root Mean Square Error of Approximation

The first value to look for unidimensionality in CFA analysis is the Chi-square (χ^2) value, and the *p*-value calculated for this value should be greater than .05 (Hooper, Coughlan and Mullen, 2008). However, they also stated that the Chi-square (χ^2) test stipulates multivariate normality as a hypothesis, and deviations from multivariate normality would result in the rejection of even a properly defined model. As can be seen in Table 4, the *p*-values calculated for the Chi-square (χ^2) values for the CFA models for each booklet used in both exams are significantly less than .05 or even .001, due to the sample size (e.g., TIMSS 2015 Booklet 1, $\chi^2(527) = 917.060$, $p<.001$). In this case, other model fit indices (goodness-of-fit) should be examined, for instance CFI, TLI and RMSEA (Hooper et al., 2008). Hair, Black, Babin and Anderson (2010) stated that indices such as CFI should be greater than .90, especially in large samples. Looking at Table 4, it can be seen that the CFI and TLI values for Booklets 1, 7, and 11 used in TIMSS 2015 are less than this cut-off value, while for the others, it is reasonable. For TIMSS 2019, on the other hand, it is seen that CFI values are greater than this cut-off value in all booklets. CFI index is actually a revised version of the TLI index (Hooper et al., 2008). On the other hand, Hair et al. (2010) stated that this cutoff value is especially valid for very complex models. In the models of this study, it is only concerned whether the items are under a single dimension. Therefore, if we take a last look at

the RMSEA value, Hooper et al. (2008) stated the cut-off value as .08. At the same time, they stated that the RMSEA value is one of the most informative fit indices due to its sensitivity to the estimated parameters in the modeling. As a matter of fact, the RMSEA values for all the booklets in the TIMSS 2015 (especially Booklets 1, 7, and 11, which cannot be unidimensional according to other indices) and the ones in the TIMSS 2019 are less than this value. As a result, we can say that the unidimensionality of the items used in all 14 booklets used in the TIMSS 2015 and TIMSS 2019 exams is at an acceptable level for each booklet.

Results of the Differential Item Functioning (DIF) Analysis

After examining the unidimensionality assumption required for DIF analysis, it can be moved on to the main analysis, DIF. The two parameters used in the DIF procedure to determine whether an item indicates bias are the beta estimate and the standardized *p*-difference index. Embretson and Reise (2013) explained the procedure for calculating the beta estimation, in which the item difficulty parameter for both groups is first calculated for each item and adjusted by subtracting the mean difference from each in the focal group. Next, the difference between the reference and adjusted focal group item difficulty parameters are calculated to have an estimate of beta for each item. Therefore, these beta estimates in SIBTEST procedure indicate whether items prefer the reference group or the focal group. Positive beta predictive values indicate item bias (DIF) in favor of the reference group (female students) while negative values indicate bias in favor of the focal group (male students). On the other hand, standardized *p*-values are calculated by looking at the difference in total scores between the reference and focal groups for each item and weighting these differences according to the focus ratios in each of the total scores. These *p*-values are then compared with an alpha level of .05 to indicate whether there is a DIF-flagged item. If the *p*-value obtained is less than .05, this item is said to indicate bias (DIF) and the null hypothesis is rejected, that is, the reference and focal groups are not equally likely to answer this item correctly. However, if the *p*-value is greater than .05, the null hypothesis is accepted and it is concluded that this item does not indicate bias (Shealy and Stout, 1993). Table 6 shows the items indicating DIF in the TIMSS 2015 and Table 7 shows the above-mentioned parameters for these items in the TIMSS 2019.

Table 6. *TIMSS 2015 mathematics items indicating gender-based item bias (DIF) at the eighth-grade level in Turkey.*

TIMSS Items	Booklet	Beta Estimate	Standard Error	χ^2	*p*-value
M042182	1	-0.1575	0.0643	5.9984	0.0143*
M042240	1	0.1226	0.0586	4.3721	0.0365*
M042164	1	0.1948	0.0674	8.3536	0.0038**
M062202	1	0.1753	0.0668	6.8925	0.0087**
M062115	3	-0.1172	0.0515	5.1740	0.0229*
M042023	5	0.2418	0.0893	7.3332	0.0068**
M042015	7	0.2041	0.0642	10.1057	0.0015**
M042114B	7	-0.1339	0.0619	4.6817	0.0305*
M042074A	7	0.1471	0.0701	4.4011	0.0359*
M042261	7	0.1764	0.0672	6.8916	0.0087**
M062244	7	0.2273	0.0741	9.4015	0.0022**
M062300	7	0.1605	0.0677	5.6195	0.0178*
M052413	9	0.1086	0.0539	4.0590	0.0439*
M052134	9	-0.1430	0.0586	5.9676	0.0146*
M062150	9	-0.1532	0.0538	8.1121	0.0044**
M062335	9	0.1377	0.0656	4.4022	0.0359*
M062133	9	0.1094	0.0557	3.8635	0.0493*
M052215	11	0.2160	0.0542	15.8642	0.0001***
M052067	11	0.1171	0.0596	3.8623	0.0494*
M062320	11	0.1565	0.0498	9.8833	0.0017**
M052125	13	0.1922	0.0529	13.1747	0.0003***
M052229	13	0.1552	0.0712	4.7576	0.0292*
M052063	13	0.1508	0.0557	7.3187	0.0068**
M052161	13	0.1489	0.0639	5.4381	0.0197*
M062192	13	0.1576	0.0489	10.3780	0.0013**

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

As seen in Table 6, according to the Differential Item Functioning (DIF) analysis performed on the eighth-grade mathematics items in TIMSS 2015, 25 out of 224 items were found to be biased (containing DIF). Only five of them express bias in favor of the focal group (male students) while the remaining 20 in favor of the reference group (female students). This means that only 2.2% of the eighth-grade mathematics items in TIMSS 2015 are biased in favor of male students, and 8.9% are biased in favor of female students. Thus, 11.1% of all items in total express item bias (DIF). Similarly, the results of the DIF analysis on the TIMSS 2019 exam are given below.

Table 7. *TIMSS 2019 mathematics items indicating gender-based item bias (DIF) at the eighth-grade level in Turkey.*

TIMSS Items	Booklet	Beta Estimate	Standard Error	χ^2	*p*-value
ME52125	1	0.3172	0.1577	4.0458	0.0443*
ME52229	1	0.4575	0.2318	3.8951	0.0484*
ME52146A	1	0.3286	0.1458	5.0783	0.0242*
ME72178C	3	0.3293	0.1212	7.3884	0.0066**
ME72027	3	-0.3409	0.1198	8.1005	0.0044**
ME52502B	5	0.2924	0.1249	5.4755	0.0193*
ME62345AB	9	0.2600	0.1250	4.3286	0.0375*
ME62171	11	0.2931	0.0988	8.8074	0.0030**
ME72221	11	-0.2315	0.0814	8.0892	0.0045**
ME62341	13	-0.2324	0.1068	4.7369	0.0295*
ME62242	13	-0.2468	0.1007	6.0088	0.0142*
ME72140D	13	0.2907	0.1287	5.1064	0.0238*
ME72154	13	-0.2860	0.1198	5.6974	0.0170*
ME72192	13	0.3018	0.1296	5.4232	0.0199*
ME72161	13	-0.3707	0.1415	6.8633	0.0088**

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

As seen in Table 7, according to the Differential Item Functioning (DIF) analysis performed on the eighth-grade mathematics items in TIMSS 2019, it was determined that 15 out of 259 items were biased (containing DIF). While six of them express bias in favor of the focal group (male students), the remaining nine in favor of the reference group (female students). This means that 2.3% of the eighth-grade mathematics items in TIMSS 2019 are biased in favor of male students while 3.5% in favor of female students. This indicates that, in total, 5.8% of all items express item bias (DIF).

Thus, the answer to the first research question of this study has emerged. In order to find an answer for the second research question, the distribution of the items with item bias in both exams according to the content and cognitive domains defined in the TIMSS conceptual framework is given below. Figure 2 shows the distribution of biased items in TIMSS 2015 and TIMSS 2019 according to content domains while Figure 3 represents according to the cognitive domains. In both figures, circular regions are created in proportion to the number of items in the specified content or cognitive domain. In this way, the number of items that indicate bias according to content domains can be compared simultaneously in terms of both the gender of the students and the year of the assessment. The number of items in each region is also indicated in the outermost circles.

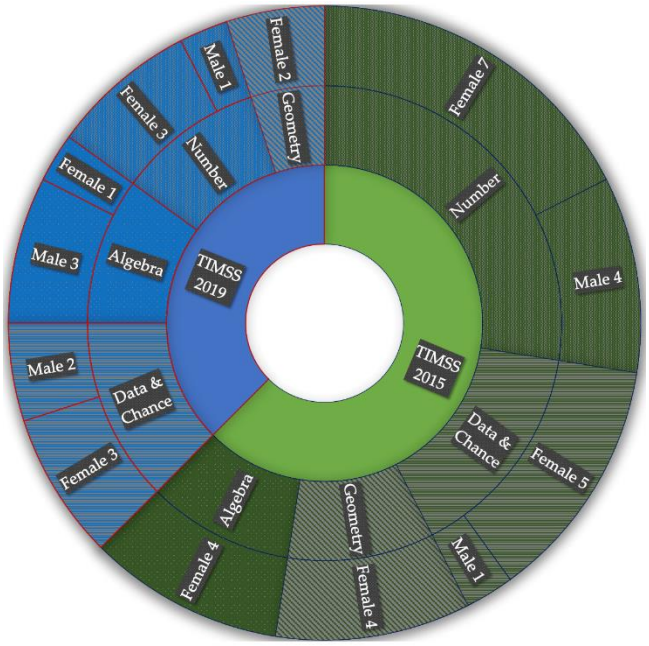


Figure 2. Distribution of eighth-grade mathematics items in favor of male and female students in TIMSS 2015 and 2019 in Turkey according to the content domains.

As seen in Figure 2, the number of items indicating item bias by gender in the TIMSS 2019 decreased remarkably when comparing with the ones in TIMSS 2015. When looking at the content domains in detail, as can be seen in the figure, there are no items indicating bias in favor of male students in either algebra or geometry content domains in TIMSS 2015. In the same year, the number of items indicating bias in favor of female students in the content domains of numbers and data and chance were relatively higher for male students. When it comes to 2019, a balance was observed in the number of items indicating bias in favor of male and female students. It may be possible to say that the balance here is due to the noticeable decrease in the number of items indicating bias in favor of female students. As stated above, there is no item indicating bias in favor of male students in the geometry content domain. In the algebra content domain, four items indicated bias in favor of female students in 2015, while there was only one item in 2019. On the other hand, while there were no items indicating bias in favor of male students in this content domain in 2015, three items indicated bias in favor of male students in 2019. Considering that only six items indicated bias in favor of male students in 2019, this is an important change. In summary, while the geometry content domain stands out with not stating bias in favor of male students in both exams; the algebra content domain, on the other hand, did not indicate a bias in favor of male students in 2015, but in 2019, it is observed as a domain that indicates more bias in favor of male students than female students. Figure 3 shows the distribution of the same items, this time in terms of cognitive domains.

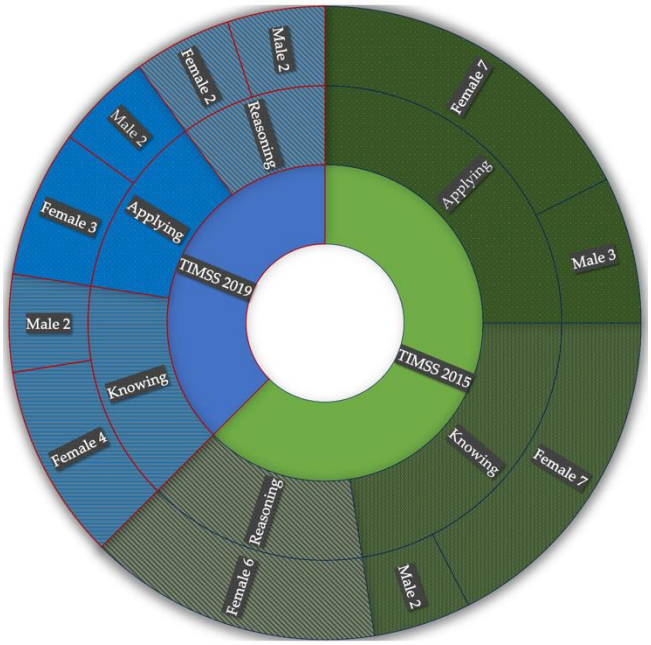


Figure 3. Distribution of eighth-grade mathematics items in favor of girls and boys in TIMSS 2015 and 2019 in Turkey according to cognitive domains.

As can be seen in Figure 3, when the distribution of items indicating item bias in the TIMSS 2015 exam is examined according to cognitive domains, it is possible to talk about a relative weight of the items indicating bias in favor of female students than male students. There is no item indicating bias in favor of male students, especially in the reasoning cognitive domain. When it comes to 2019, it is seen that items expressing bias in cognitive domains, similar to content domains, show a balanced distribution between genders. Here, there is no significant change in the items indicating bias in favor of male students. This balanced situation that emerged in 2019 is the significant decrease in the number of items indicating bias in favor of female students. In addition to these figures, Appendix 1 and Appendix 2, in addition to the distribution of biased items in both assessments according to content and cognitive domains, illustrate the content topics and item labels of the DIF-flagged items. It can be an important resource for researchers who want to work on this subject.

Discussion, Conclusion and Recommendations

In terms of mathematics achievement, from 2015 to 2019, Turkey increased its place from 29th place among 46 participants (Mullis et al., 2016) to 24th place among 46 participants (Mullis et al., 2020) in the TIMSS in the field of mathematics at the eighth-grade level. While an increase in success is observed, there is no significant difference between male and female students in terms of overall mean achievement scores in both assessments (Table 2). This situation does not coincide with the findings of studies on Turkey in terms of PISA 2006 (Alacacı and Erbaş, 2010; Demir et al., 2010; Dinçer and Kolasin, 2009) and PISA 2009 results (Gürsakal, 2012). In the mentioned studies, it was observed that female students were significantly more successful than male students. On the other hand, according to a study on TIMSS 2007 results for Turkey, gender was stated as a neutral indicator in terms of student achievement (Atar, 2011). Similarly, no significant difference in achievement was observed between

male and female students in studies conducted by creating their own samples in schools instead of using international data (Aksu, 2001; Işıksal and Aşkar, 2005). As a result, unlike the PISA results for Turkey, there is no significant difference in mathematics performance between male and female students in the TIMSS 2015 and 2019 assessments. However, as stated in Table 2, when looking at the TIMSS 2015 and 2019 results in terms of content and cognitive domains, it can be observed that there are significant differences between male and female students both in the overall participant countries and in Turkey. Although there is no difference in general terms, these differences in terms of content and cognitive domains may support the bias due to the structure of the items.

As a result of this study carried out from this point of view, according to the Item Functioning Analysis (DIF) conducted on the eighth-grade mathematics items in the TIMSS 2015, 25 out of 224 items were found to be biased (containing DIF). Only five of them express bias in favor of the focal group (male students) while the remaining 20 in favor of the reference group (female students). In addition, in the eighth-grade mathematics items in TIMSS 2019, 15 out of 259 items were found to be biased (containing DIF). Six of them express bias in favor of the focal group (male students) while the remaining nine in favor of the reference group (female students). In this study, the SIBTEST procedure, which was created according to the Item Response Theory, was used rather than the methods created according to the Classical Test Theory while performing the DIF analysis. As stated in previous studies (Hambleton and Jones, 1993; Kan et al., 2013; Karakaya, 2012; Kelecioğlu et al., 2014), it should be noted that more sensitive DIF measurements could be made in this way. As a result, 11.1% of the items in the TIMSS 2015 (2.2% in favor of male students, 8.9% in favor of female students) and 5.8% of the items in the TIMSS 2019 (2.3% in favor of male students, 3.5% in favor of female students) stated bias in genders. From the point of view of male students, there is no significant change in the percentage of items that indicate bias, while the situation appears as a significant decrease in terms of female students. However, in 2019, the number of items indicating bias in favor of female students is still higher than that of male students. Thus, the necessity of examining these items in terms of content and cognitive domains has emerged once again.

As a result of the DIF analysis conducted in this study, in addition to whether the mathematics items used in the TIMSS 2015 and TIMSS 2019 indicate a bias in favor of eighth-grade female or male students in Turkey, the distribution of the biased items in terms of content and cognitive domains specified in the TIMSS conceptual framework was also examined (Figure 2). According to the analysis made in terms of content domains,

– the geometry content domain stands out with the fact that it does not indicate bias in favor of male students in both exams,

– While the content domain of algebra did not indicate a bias in favor of male students in 2015, it is observed as a domain that indicates more bias in favor of male students than female students in 2019.

The first thing that comes to mind here is the DIF analysis study conducted by Bakan-Kalaycıoğlu and Kelecioğlu (2011) on the ÖSS exam (college placement test). In their study, they identified bias in three of the 45 mathematics items used in the 2005 ÖSS exam. They found that two geometry items indicated a bias in favor of male students and one algebra item in favor of female students. In addition, studies conducted outside of Turkey in this field also indicate that male students are more successful in geometry and female students in algebra than the opposite sex (Lane et al., 1996; McGraw et al., 2006). 43 of the 224 items used in the TIMSS 2015 exam and 54 of the 259 items in the TIMSS 2019 belong to the geometry content domain (Figure 1). Interestingly, none of the geometry items used in these two assessments indicate bias in favor of male students. Similarly, 64 items in TIMSS 2015 and 69 items in TIMSS 2019 belong to the algebra content domain. While there were no algebra items in favor of male students in 2015, this situation was reversed in 2019, and three of the four algebra items indicating DIF indicate a bias in favor of male students. In this sense, it is seen that the findings of the current study on TIMSS do not match with the findings of the study conducted on the college placement test of Bakan-Kalaycıoğlu and Kelecioğlu (2011) and the findings of studies conducted outside of Turkey (Lane et al., 1996; McGraw et al., 2006). The fact that the findings of this study do not overlap with the studies conducted with samples from both Turkey and abroad reveals that the items used in the TIMSS assessment can affect eighth-grade female and male students in different ways in Turkey.

On the other hand, the items used in the TIMSS 2015 and TIMSS 2019 were also examined in terms of cognitive domains (Figure 3). As a result,

– It is possible to talk about a relative weight of the items indicating bias in favor of female students for male students in TIMSS 2015. There is no item indicating bias in favor of male students, especially in the reasoning cognitive domain.

– When it comes to 2019, it is seen that items expressing bias in cognitive domains, similar to content domains, show a balanced distribution between genders.

Here, the reasoning domain came to the forefront with no items in favor of male students in 2015. This situation is in line with the findings of previous studies, unlike the situation in the content domains. Because it has been revealed by previous researchers that female students are relatively more successful than male students in reasoning problems (Friedman, 1996; Ryan and Chiu, 1996). In addition, in 2015, no significant difference was observed between male and female students in the TIMSS in all three cognitive domains (Table 2). However, in 2019, it is statistically seen that female students have significantly higher mean scores than male students in the cognitive domains of knowing

and reasoning. On the other hand, while there were more biased items in favor of female students in terms of cognitive domains in TIMSS 2015, this situation was balanced in 2019.

Whether it is TIMSS 2015 or TIMSS 2019, unfortunately, the contents of the items used in these assessments are not completely shared as open data. In this respect, especially in line with the findings of this study, the content domains of geometry and algebra and the cognitive domain of reasoning should be discussed in more detail in terms of the bias between male and female students in Turkey. Further studies that may examine these domains in terms of the inclusion in the textbooks concerning gender as well as the ones that may focus on the attitudes of male and female students towards these domains would contribute to the literature. As a result of the studies carried out on samples, especially in Turkey and abroad, it is observed that the geometry content domain is known for the success of male students and the algebra content domain for female students. Similarly, these domains were also highlighted in regard to the findings of the previous studies as favoring the indicated gender groups in this study, while the results of TIMSS 2015 and TIMSS 2019 show exactly the opposite of this situation. In this sense, further studies that examine how these domains are handled in the textbooks, in the classroom instruction, etc. in terms of the genders in detail would be helpful in the continuation of this current study. Appendix 1 and Appendix 2, in addition to the distribution of the items that were concluded to be biased in this study in terms of content and cognitive domains, illustrate the content topics and item labels of these items. Researchers can use it without hesitation in terms of contributing to the studies to be done on this subject.

References

Aaronson, D., Barrow, L. & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics*, *25*(1), 95-135.

Aksu, M. (2001). Student performance in dealing with fractions. *The Journal of Educational Research, 90*(6), 375-380.

Akyüz, G. (2006). Investigation of the effect of teacher and class characteristics on mathematics achievement in Turkey and European Union countries. *Elementary Education Online*, *5*(2), 75-86.

Akyüz, G. (2014). The effects of student and school factors on mathematics achievement in TIMSS 2011. *Egitim ve Bilim, 39*(172), 150-162.

Akyüz, G. & Berberoğlu, G. (2010). Teacher and classroom characteristics and their relations to mathematics achievement of the students in the TIMSS. *New Horizons in Education, 58*(1), 77-95.

Alacacı, C. & Erbaş, A. K. (2010). Unpacking the inequality among Turkish schools: Findings from PISA 2006. *International Journal of Educational Development*, *30*(2), 182-192.

Atar, B. (2011). Application of descriptive and explanatory item response models to TIMSS 2007 Turkey mathematics data. *Egitim ve Bilim*, *36*(159), 255-269.

Awuor, R. A. (2008). *Effect of unequal sample sizes on the power of DIF detection: An IRT-based Monte Carlo study with SIBTEST and Mantel-Haenszel procedures.* Unpublished Doctoral dissertation, Virginia Polytechnic Institute and State University.

Bakan-Kalaycıoğlu, D. & Kelecioğlu, H. (2011). Öğrenci seçme sınavı'nın madde yanlılığı açısından incelenmesi. *Eğitim ve Bilim, 36*(161), 3-13.

Bilican, S., Demirtaşlı, R. N. & Kilmen, S. (2011). The attitudes and opinions of the students towards mathematics course: The comparison of TIMSS 1999 and TIMSS 2007. *Educational Sciences: Theory and Practice*, *11*(3), 1277-1283.

Boyd, D., Grossman, P., Lankford, H., Loeb, S. & Wyckoff, J. (2005). How changes in entry requirements alter the teacher workforce and affect student achievement. (Working Paper No. 11844). Cambridge, MA: National Bureau of Economic Research.

Clotfelter, C. T., Ladd, H. F. & Vigdor, J. L. (2007). Teacher credentials and student achievement: Longitudinal analysis with student fixed effects. *Economics of Education Review*, *26*(6), 673-682.

Clotfelter, C. T., Ladd, H. F. & Vigdor, J. L. (2010). Teacher credentials and student achievement in high school a cross-subject analysis with student fixed effects. *Journal of Human Resources*, *45*(3), 655-681.

Demir, I., Kılıç, S. & Ünal, H. (2010). Effects of students' and schools' characteristics on mathematics achievement: Findings from PISA 2006. *Procedia Social and Behavioral Sciences, 2*(2010), 3099-3103.

Dinçer, M. A. & Kolasin, G. U. (2009). *Türkiye'de öğrenci başarısında eşitsizliğin belirleyicileri*. İstanbul: Sabancı Üniversitesi Eğitim Girişimi Reformu.

Doğan, N. & Barış, F. (2010). Tutum, değer ve özyeterlik değişkenlerinin TIMSS-1999 ve TIMSS-2007 sınavlarında öğrencilerin matematik başarılarını yordama düzeyleri. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *1*(1), 44-50.

Else-Quest, N. M., Hyde, J. S. & Linn, M. C. (2010). Cross-national patterns of gender differences in mathematics: A meta-analysis. *Psychological Bulletin, 136*(1), 101–127.

Embretson, S. E. & Reise, S. P. (2013). *Item response theory*. Psychology Press.

Engin-Demir, C. (2009). Factors influencing the academic achievement of the Turkish urban poor. *International Journal of Educational Development, 29*(2009), 17-29.

Erkan, S. S. S. (2013). A comparison of the education systems in Turkey and Singapore and 1999–2011 TIMSS tests results. *Procedia-Social and Behavioral Sciences*, *106*(2013), 55-64.

Friedman, L. (1996). Meta-analysis and quantitative gender differences: Reconciliation. *Focus on Learning Problems in Mathematics, 18*(3), 123-128.

Goldhaber, D. & Anthony, E. (2007). Can teacher quality be effectively assessed? National board certification as a signal of effective teaching. *The Review of Economics and Statistics*, *89*(1), 134-150.

Gronmo, L. S., Linquist, M., Arora, A. & Mullis, I. V. S. (2013). TIMSS 2015 mathematics framework. In Mullis, I. V. S., Martin, M. O. (Eds.). *TIMSS 2015 assessment frameworks (pp. 11-27).* Chestnut Hill, MA: TIMSS & PIRLS International Study Center at Boston College.

Güner, N., Sezer, R. & İspir, O. A. (2013). İlköğretim ikinci kademe öğretmenlerinin TIMSS hakkındaki görüşleri. *Pamukkale Üniversitesi Eğitim Fakültesi Dergisi, 33*(1), 11-29.

Gürsakal, S. (2012). An evaluation of PISA 2009 student achievement levels' affecting factors. *Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Dergisi, 17*(1), 441-452.

Hair, J. F., Black, W. C., Babin, B. J. & Anderson, R. (2010). *Multivariate analysis (7th ed.)*. Pearson Prentice Hall.

Hambleton, R. K. & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, *12*(3), 38-47.

Hill, H. C., Rowan, B. & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal*, *42*(2), 371-406.

Hooper, D., Coughlan, J. & Mullen, M. R. (2008). Equation modelling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods*, *6*(1), 53-60.

İncikabı, L. (2012). After the reform in Turkey: A content analysis of SBS and TIMSS assessment in terms of mathematics content, cognitive domains, and item types. *Education as Change*, *16*(2), 301-312.

Işıksal, M. & Aşkar, P. (2005). The effect of spreadsheet and dynamic geometry software on the achievement and self-efficacy of 7th-grade students. *Educational Research, 47*(3), 333-350.

Jacob, B. A. & Lefgren, L. (2002). The impact of teacher training on student achievement: Quasi-experimental evidence from school reform efforts in Chicago. (Working Paper No. 8916). Cambridge, MA: The National Bureau of Economic Research.

Jöreskog, K. G. & Sörbom, D. (1996). *LISREL 8: User's reference guide*. Scientific Software International.

Kan, A., Sünbül, Ö. & Ömür, S. (2013). 6.-8. sınıf seviye belirleme sınavları alt testlerinin çeşitli yöntemlere göre değişen madde fonksiyonlarının incelenmesi. *Mersin Üniversitesi Eğitim Fakültesi Dergisi*, *9*(2), 207-222.

Karakaya, I. (2012). An investigation of item bias in science and technology subtests and mathematic subtests in level determination exam (LDE). *Educational Sciences: Theory and Practice*, *12*(1), 222-229.

Kelecioğlu, H., Karabay, B. & Karabay, E. (2014). Investigation of placement test in terms of item biasness. *Elementary Education Online*, *13*(3), 934-953.

Kılıç, S., Çene, E. & Demir, I. (2012). Comparison of learning strategies for mathematics achievement in Turkey with eight countries. *Educational Sciences, 12*(4), 2594-2598.

Lane, S., Wang, N. & Magone, M. (1996). Gender-related differential item functioning on a middle-school mathematics performance assessment. *Educational Measurement: Issues and Practice, 15*(4), 21-27.

Magis, D., Beland, S., Tuerlinckx, F. & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Rsearch Methods*, *42*(3), 847-862.

Martin, M. O., Mullis, I. V. & Foy, P. (2008). *TIMSS 2007 international mathematics report: Findings from IEA's Trends in International Mathematics and Science Study at the fourth and eighth grades*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center at Boston College.

Martin, M. O., Mullis, I. V. S. & Foy, P. (2013). TIMSS 2015 assessment design. In Mullis, I. V. S., Martin, M. O. (Eds.). *TIMSS 2015 assessment frameworks (pp. 85-99).* Chestnut Hill, MA: TIMSS & PIRLS International Study Center at Boston College.

Martin, M. O., Mullis, I. V. S. & Foy, P. (2017). TIMSS 2019 assessment design. In Mullis, I. V. S., & Martin, M. O. (Eds.). *TIMSS 2019 assessment frameworks (pp. 81-91).* Chestnut Hill, MA: TIMSS & PIRLS International Study Center at Boston College.

McGraw, R., Lubienski, S. T. & Strutchens, M. E. (2006). A closer look at gender in NAEP mathematics achievement and affect data: Intersections with achievement, race/ethnicity, and socioeconomic status. *Journal for Research in Mathematics Education, 37*(2), 129-150.

Mullis, I. V., Martin, M. O., Foy, P. & Arora, A. (2012). *TIMSS 2011 international results in mathematics*. International Association for the Evaluation of Educational Achievement. Chestnut Hill, MA: TIMSS & PIRLS International Study Center at Boston College.

Mullis, I. V. S., Martin, M. O., Foy, P. & Hooper, M. (2016). *TIMSS 2015 international results in mathematics.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center at Boston College.

Mullis, I. V. S., Martin, M. O., Foy, P., Kelly, D. L. & Fishbein, B. (2020). *TIMSS 2019 international results in mathematics and science.* Retrieved from: https://timssandpirls.bc.edu/timss2019/international-results/

National Center for Education Statistics. (2021). *International data explorer*. Retrieved from: https://nces.ed.gov/timss/idetimss/

Nye, B., Konstantopoulos, S. & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis*, *26*(3), 237-257.

Osterlind, S. J. & Everson, H. T. (2009). *Differential item functioning*. Sage Publications.

Rosseel, Y. (2012). "lavaan: An R Package for structural equation modeling." *Journal of Statistical Software, 48*(2), 1–36. Retrieved from: https://www.jstatsoft.org/v48/i02/.

Rowan, B., Chiang, F. S. & Miller, R. J. (1997). Using research on employees' performance to study the effects of teachers on students' achievement. *Sociology of Education*, *70*(4), 256-84.

Ryan, K. E. & Chiu, S. (1996). *Detecting DIF on mathematics items: The case for gender and calculator sensitivity.* Paper presented at the annual meeting of the American Education Research Association, New York, NY.

Shealy, R. & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika*, *58*(2), 159-194.

Stout, W. & Roussos, L. (1996). DIF-pack SIBTEST program [Open source computer software].

Stronge, J. H., Ward, T. J. & Grant, L. W. (2011). What makes good teachers good? A cross-case analysis of the connection between teacher effectiveness and student achievement. *Journal of Teacher Education*, *62*(4), 339-355.

Yıldırım, H. H., Yıldırım, S., Ceylan, E., Yetişir, M. İ. & Ajans, C. (2013). *Türkiye perspektifinden TIMSS 2011 sonuçları*. Pelin Ofset: Ankara, Turkey.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and likert-type (ordinal) item scores.* Ottawa,

ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Ek 1. TIMSS 2015 Sınavında Türkiye'de Kız ve Erkek Öğrencilerin Lehine Yanlılık Gösteren 8. Sınıf Matematik Sorularının Dağılımları

Odak Grubu (Erkek Öğrenciler) Lehine Yanlılık Belirten Sorular				
Soru	*İçerik Alanı*	*Bilişsel Alan*	*İçerik Konusu*	*Soru Başlığı*
M052134	Sayılar	Bilme	Kesirler, Ondalıklı ve Tam Sayılar	A şehri B şehrinden ne kadar sıcak
M062150	Sayılar	Bilme	Kesirler, Ondalıklı ve Tam Sayılar	X ve Y şehirleri arasındaki düşük sıcaklık farkı
M042182	Sayılar	Uygulama	Kesirler, Ondalıklı ve Tam Sayılar	John ne kadar yükseğe sıçradı?
M042114B	Sayılar	Uygulama	Oran, Orantı ve Yüzde	28 mm'lik bir yığındaki kağıt sayısı
M062115	Veri & Şans	Uygulama	Şans	Mermer labirentte kesişimler
Referans Grubu (Kız Öğrenciler) Lehine Yanlılık Belirten Sorular				
Soru	*İçerik Alanı*	*Bilişsel Alan*	*İçerik Konusu*	*Soru Başlığı*
M052413	Sayılar	Bilme	Kesirler, Ondalıklı ve Tam Sayılar	Verilen sayısal ifadeyi çözmek
M052229	Sayılar	Bilme	Kesirler, Ondalıklı ve Tam Sayılar	Ondalıklı sayıyı kesre dönüştürme
M052215	Sayılar	Bilme	Kesirler, Ondalıklı ve Tam Sayılar	Taralı bölgenin kesir olarak ifadesi
M062335	Sayılar	Bilme	Oran, Orantı ve Yüzde	3:2'ye eşdeğer olan oranı seçmek
M042015	Sayılar	Bilme	Doğal Sayılar	3 sayısının küpünün değeri
M042023	Sayılar	Uygulama	Kesirler, Ondalıklı ve Tam Sayılar	7/12 mi 2/3'mü daha büyüktür?
M052125	Sayılar	Akıl Yürütme	Doğal Sayılar	3'ün katları
M052067	Cebir	Bilme	Cebirsel İfadeler ve İşlem	Verilen cebirsel ifadenin değeri
M042240	Cebir	Uygulama	Cebirsel İfadeler ve İşlem	X ve Y'nin değeri
M052063	Cebir	Uygulama	Cebirsel İfadeler ve İşlem	Dikdörtgenin alanı için cebirsel ifade
M042074A	Cebir	Akıl Yürütme	Fonksiyonlar	4 ve 30 desenleri için daireler
M062244	Geometri	Uygulama	Geometrik Dönüşümler	AB doğrusunun orta noktasının koordinatları
M062300	Geometri	Akıl Yürütme	Geometride Ölçme	Alan ve çevresi verilen bir dikdörtgen çizmek
M062202	Geometri	Akıl Yürütme	Geometrik Şekiller	Liza'nın küpün açılımı gösterimi – Q yüzünün karşıt yüzü
M062192	Geometri	Akıl Yürütme	Geometrik Şekiller	Merdiven tabanından bina tabanına olan mesafe
M042261	Veri & Şans	Bilme	Şans	Yağmur yağma ihtimali
M062133	Veri & Şans	Uygulama	Şans	Siyah ve beyaz mermerler olan torbadan değiştirerek çekiliş
M052161	Veri & Şans	Uygulama	Şans	Bir torbadaki topların sayısı
M062320	Veri & Şans	Uygulama	Veri Yorumlama	Balık tutma histogramı – Serbest bırakılan balıkların oranı
M042164	Veri & Şans	Akıl Yürütme	Veri Yorumlama	Tezgâhtarla aynı/farklı görüşte olmak

Ek 2. TIMSS 2019 Sınavında Türkiye'de Kız ve Erkek Öğrencilerin Lehine Yanlılık Gösteren 8. Sınıf Matematik Sorularının Dağılımları

Odak Grubu (Erkek Öğrenciler) Lehine Yanlılık Belirten Sorular				
Soru	*İçerik Alanı*	*Bilişsel Alan*	*İçerik Konusu*	*Soru Başlığı*
ME72027	Sayılar	Uygulama	Kesirler ve Ondalıklı Sayılar	Dolu kaptaki su oranı
ME72221	Cebir	Bilme	Cebirsel İfadeler ve İşlem	Bilinen x değeri için $y = 2x^2 - 2x + 5$ ifadesinin değeri
ME62341	Cebir	Bilme	Fonksiyonlar	y=3x-1 doğrusunun eğimi
ME62242	Cebir	Akıl Yürütme	Fonksiyonlar	Doğrusal ilişki
ME72154	Veri & Olasılık	Uygulama	Veri	Dünyadaki internet kullanıcılarının sütun grafiği
ME72161	Veri & Olasılık	Akıl Yürütme	Veri	Basketbol takımının ortalama boyu
Referans Grubu (Kız Öğrenciler) Lehine Yanlılık Belirten Sorular				
Soru	*İçerik Alanı*	*Bilişsel Alan*	*İçerik Konusu*	*Soru Başlığı*
ME52229	Sayılar	Bilme	Kesirler ve Ondalıklı Sayılar	Ondalıklı sayıyı kesre dönüştürme
ME72178C	Sayılar	Bilme	Tam Sayılar	Tam kare (81) ile etiketlenmiş raflar
ME52125	Sayılar	Akıl Yürütme	Tam Sayılar	3'ün katları
ME52146A	Cebir	Akıl Yürütme	Fonksiyonlar	Bir şekil için gerekli kibrit sayısı
ME62171	Geometri	Bilme	Geometrik Şekiller ve Ölçme	Taralı alanın XY doğrusu üzerinde yansıması
ME72140D	Geometri	Bilme	Geometrik Şekiller ve Ölçme	Şeklin doğru üzerinden yansıması
ME52502B	Veri & Olasılık	Uygulama	Veri	Bir ürün için en düşük fiyat veren mağaza
ME62345AB	Veri & Olasılık	Uygulama	Veri	Ortalama ve medyan değerleri ile balık tutma noktası belirleme
ME72192	Veri & Olasılık	Uygulama	Veri	En sevilen meyveyi en iyi gösteren grafik

Appendix 1. Distribution of eighth-grade mathematics items favoring genders in Turkey in TIMSS 2015

Items Favoring Focal Group (Male Students)				
Item	*Content Domain*	*Cognitive Domain*	*Topic Area*	*Item Label*
M052134	Number	Knowing	Fractions, Decimals, and Integers	How much hotter is city A than B
M062150	Number	Knowing	Fractions, Decimals, and Integers	Difference between low temperature in City X and Y
M042182	Number	Applying	Fractions, Decimals, and Integers	How far did john jump
M042114B	Number	Applying	Ratio, Proportion, and Percent	Number of papers in a 28mm stack
M062115	Data & Chance	Applying	Chance	Marble maze with intersections
Items Favoring Reference Group (Female Students)				
Item	*Content Domain*	*Cognitive Domain*	*Topic Area*	*Item Label*
M052413	Number	Knowing	Fractions, Decimals, and Integers	Solve given numeric expression
M052229	Number	Knowing	Fractions, Decimals, and Integers	Convert decimal to a fraction.
M052215	Number	Knowing	Fractions, Decimals, and Integers	Fraction of diagram shaded
M062335	Number	Knowing	Ratio, Proportion, and Percent	Select equivalent ratio to 3:2
M042015	Number	Knowing	Whole Numbers	What is the value of cube of 3
M042023	Number	Applying	Fractions, Decimals, and Integers	Which is larger 7/12 or 2/3
M052125	Number	Reasoning	Whole Numbers	Multiples of 3
M052067	Algebra	Knowing	Expressions and Operations	Value of given expression
M042240	Algebra	Applying	Expressions and Operations	Value of x and y
M052063	Algebra	Applying	Expressions and Operations	Expression for area of rectangle
M042074A	Algebra	Reasoning	Relationships and Functions	Circles for patterns 4 & 30
M062244	Geometry	Applying	Location and Movement	Find coordinates of midpoint of line AB
M062300	Geometry	Reasoning	Geometric Measurement	Draw a rectangle given area and perimeter
M062202	Geometry	Reasoning	Geometric Shapes	Liza's net of cube - face opposite face Q
M062192	Geometry	Reasoning	Geometric Shapes	Distance from base of ladder to base of building
M042261	Data & Chance	Knowing	Chance	How likely it will rain
M062133	Data & Chance	Applying	Chance	Black and white marbles in a bag with replacement
M052161	Data & Chance	Applying	Chance	Number of balls in a bag
M062320	Data & Chance	Applying	Data Interpretation	Fishing histogram - proportion of fish released
M042164	Data & Chance	Reasoning	Data Interpretation	Agree/disagree with the salesman

Appendix 2. Distribution of eighth-grade mathematics items favoring genders in Turkey in TIMSS 2019

Items Favoring Focal Group (Male Students)				
Item	*Content Domain*	*Cognitive Domain*	*Topic Area*	*Item Label*
ME72027	Number	Applying	Fractions and Decimals	Fraction of water in full container
ME72221	Algebra	Knowing	Expressions and Operations	Value of $y = 2x^2 - 2x + 5$ given x
ME62341	Algebra	Knowing	Relationships and Functions	Slope of line $y = 3x - 1$
ME62242	Algebra	Reasoning	Relationships and Functions	Relationship for a linear graph in words
ME72154	Data & Chance	Applying	Data	Bar graph of Internet users in world
ME72161	Data & Chance	Reasoning	Data	Mean heights of basketball team
Items Favoring Reference Group (Female Students)				
Item	*Content Domain*	*Cognitive Domain*	*Topic Area*	*Item Label*
ME52229	Number	Knowing	Fractions and Decimals	Convert decimal to a fraction
ME72178C	Number	Knowing	Integers	Shelves labeled with perfect square - 81
ME52125	Number	Reasoning	Integers	Multiples of 3
ME52146A	Algebra	Reasoning	Relationships and Functions	Number of matches for figure 10
ME62171	Geometry	Knowing	Geometric Shapes and Measurement	Shaded square reflected over the line XY
ME72140D	Geometry	Knowing	Geometric Shapes and Measurement	Shape reflected over dotted line - D
ME52502B	Data & Chance	Applying	Data	Store with lowest price per pair - S
ME62345AB	Data & Chance	Applying	Data	Fishing spots - calculate mean and median - LB Median
ME72192	Data & Chance	Applying	Data	Graph that best shows favorite fruit