Research Article / **Araştırma Makalesi**

Bolu Abant İzzet Baysal Üniversitesi

Eğitim Fakültesi Dergisi (BAİBÜEFD)

Bolu Abant Izzet Baysal University

Journal of Faculty of Education

2025, 25(1), 398– 424. https://doi.org/10.17240/aibuefd.2025.25.90529-1441270

Angoff'un Dönüştürülmüş Madde Güçlükleri Yöntemi'nin Değişen Madde Fonksiyonu Belirlemede Kullanımı*[1]

The Use of Angoff's Transformed Item Difficulties Method in Detecting Differential Item Functioning

Metehan GÜNGÖR[2], **Ergül DEMİR**[3]

Geliş Tarihi (*Received*): 22.02.2024 **Kabul Tarihi (*Accepted*):** 19.12.2024 **Yayın Tarihi (*Published*):** 15.03.2025

Öz: Bu çalışmada değişen madde fonksiyonu (DMF) belirleme yöntemlerinden Angoff'un Dönüştürülmüş Madde Güçlükleri (Transformed Item Difficulties) Yöntemi önemli ayrıntıları ve eleştirilen yönleriyle tanıtılmakta, yöntemin güçlü ve zayıf yönleri tartışılmaktadır. Tek boyutlu ve 0-1 şeklinde puanlanan maddelerden oluşan testlerde DMF belirleme çalışmalarında kullanılmak üzere geliştirilmiş öncü yöntemlerden biri olan Angoff'un yöntemine yönelik olarak alan yazında bazı eleştiriler bulunmaktadır. Yöntemin yalnızca madde güçlüklerine odaklı olması, gruplar arasındaki gerçek farkın değişen madde fonksiyonu olarak görülme olasılığı bulunması gibi nedenlerle bu yöntemin kullanılmaması önerilebilmektedir. Diğer taraftan yöntem, uygulama kolaylığı, grafiksel yorumlama imkânı tanıması ve görece küçük örneklemlerde de kullanılabilmesi gibi pratik avantajlara sahiptir. Bu kapsamda bu çalışmada Angoff'un yönteminin algoritması, genel karakteristiği, güçlü yönleri ve sınırlılıkları tartışılmıştır. Ayrıca, R programlama dili üzerinde kullanılabilen "difR" paketi ile DMF analizinde Angoff'un yönteminin nasıl kullanılacağı adım adım satır komutları yardımıyla açıklanmıştır. Yürütülen tartışmalar göstermektedir ki, Angoff'un Yöntemi ile DMF belirleme, dikkat edilmesi gereken bazı önemli sınırlılıklar içermektedir. Bununla birlikte yöntemin uygulama kolaylığı ve görselleştirme imkânı tanıyor oluşu, yanlılık ve DMF kavramlarının temellerinin açıklanması açısından yararlı olabilir. Bu yöntem, grupların ölçülen özellik bakımından test puanı ortalamalarının yakın ya da eşit olması durumunda daha anlamlı sonuçlar verebilmektedir. Yöntemin, bir testteki potansiyel olarak yanlı maddelerin belirlenmesinde bir öngörü sağlaması amacıyla daha sınırlı kullanımı düşünülebilir.

Anahtar Kelimeler: Değişen Madde Fonksiyonu, Dönüştürülmüş Madde Güçlükleri, Difr, Delta Yöntemi, DMF Analizi.

&

Abstract: In this study, the Angoff's Transformed Item Difficulties method of detecting differential item functioning (DIF) is introduced with its important details and criticized aspects. The strengths and weaknesses of this Method are discussed. Among the pioneering methods developed for DIF detection in tests consisting of single-dimension items scored on a 0-1 scale, there are some criticisms in the literature regarding Angoff's Method. It is suggested that this Method may not be used due to criticisms such as its exclusive focus on item difficulties and the possibility of viewing real differences between groups as item bias. On the other hand, the Method has practical advantages such as ease of application, the possibility of graphical interpretation, and usability in relatively small samples. Within the scope of this study, the algorithm, general characteristics, strengths, and limitations of Angoff's Method are discussed. In addition, the step-by-step command lines for using Angoff's Method in DIF analysis with the "difR" package in the R programming language are explained. The discussions conducted indicate that using Angoff's Method for DIF detection comes with some important limitations that need to be taken into account. However, the Method's ease of application and visualization capabilities can be beneficial in explaining the fundamentals of bias and DIF concepts. This Method can provide more meaningful results when the test score averages of groups are close or equal in terms of the measured characteristics. The Method's limited use can be considered for the purpose of identifying potentially biased items in a test.

Keywords: Differential Item Functioning, Transformed Item Difficulties, Difr, Delta Plot, DIF Analysis.

Atıf/Cite as: Güngör, M., & Demir, E. (2025). Angoff'un Dönüştürülmüş Madde Güçlükleri Yöntemi'nin değişen madde fonksiyonu belirlemede kullanımı. *Bolu Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi*, 25(1), 398-424, https://doi.org/10.17240/aibuefd.2025.25.90529-1441270

İntihal-Plagiarism/Etik-Ethic: Bu makale, en az iki hakem tarafından incelenmiş ve intihal içermediği, araştırma ve yayın etiğine uyulduğu teyit edilmiştir. / This article has been reviewed by at least two referees and it has been confirmed that it is plagiarism-free and complies with research and publication ethics. https://dergipark.org.tr/tr/pub/aibuelt

* Bu çalışmanın bir kısmı 20-23 Eylül 2023 tarihleri arasında Ankara Üniversitesi'nde düzenlenen EduCongress 2023'te sunulmuştur.

[1] Bu çalışma TÜBİTAK tarafından desteklenen 223K382 numaralı proje kapsamında gerçekleştirilmiştir.

[2] Sorumlu Yazar: Metehan GÜNGÖR, Ankara Üniversitesi, Eğitim Bilimleri Enstitüsü, mhgungor@ankara.edu.tr, ORCID: 0000-0003-4409-2229

[3] Doç. Dr. Ergül DEMİR, Ankara Üniversitesi, Eğitim Bilimleri Enstitüsü, erguldemir@ankara.edu.tr, ORCID: 0000-0002-3708-8013

1. GİRİŞ

1.1. Madde Yanlılığı ve Değişen Madde Fonksiyonu

Ölçme araçlarından alınan puanlar çeşitli amaçlarla kullanılabilmektedir ve bu puanlar bazı önemli kararlara esas oluşturmaktadır. Bu kararların yerinde olabilmesi için ölçme aracından alınan puanların geçerli olması gerekmektedir. Bu nedenle araştırmacılar/test geliştiriciler ölçme aracından alınan puanların geçerliğine ilişkin kanıtlar sunmak durumundadır. Bu kanıtların çeşitli şekillerde toplanabileceği bilinmektedir ve ölçme araçlarında geçerlik kavramı kapsamlı bir konudur. Ancak bu çalışmada, Eğitsel ve Psikolojik Test Standartları [Standards for Educational and Psychological Testing] (AERA vd., 1999) ve Eğitimde Adil Test Uygulamaları Kuralları [Code of Fair Testing Practices in Education] (APA, 1988) içerisinde önemle vurgulanan "test adilliği [fairness]" konusu odağa alındığı için özellikle test adilliği ile ilgili olan geçerlik kanıtlarına ilişkin açıklamalara ve tartışmalara yer verilmiştir.

Test geliştiriciler standart geçerlik kanıtlarını sundukları gibi test puanlarının yanlılıktan da arınık olmasını garanti altına almalıdır. Testin geneli yanlı olabileceği gibi, testteki maddeler de yanlı olabilir. Bir testin geneli ya da testte yer alan maddeler belli bir alt grup lehine ya da aleyhine yanlı davranıyor ise bu testten alınan puanlarda geçerlik sorunları ortaya çıkabilir ve bu nedenle test puanları üzerinden yapılacak karşılaştırmalar yanıltıcı olabilir. Bir ölçme aracından alınan puanların, ölçme aracını yanıtlayanların testle ölçülen özellikleri dışındaki diğer bazı özelliklerine bağlı olması durumu, test ve madde yanlılığı risklerini ortaya çıkarmaktadır. Farklı gruplardaki bireylerin bir testteki maddeleri yanıtlamada farklı yanıtlama davranışları göstermesi de madde yanlılığı çalışmalarının yürütülmesini gerektiren durumlar arasında özellikle belirtilmektedir (Mellenberg, 1989; van der Flier vd., 1984). Farklı gruplara (ırk, cinsiyet vb.) ait yanıtlayıcıların test maddelerine verdikleri yanıtlarda farklılaşma söz konusu ise bu durumun etkileri mutlaka incelenmelidir.

Bu çalışmada özellikle madde düzeyindeki yanlılık konusu ele alınmaktadır. Madde yanlılığı kısaca, testteki bir maddenin belirli bir alt grup lehine veya aleyhine çalışmasıdır (Osterlind ve Everson, 2009). Bu tür bir yanlılık testten alınan puanların geçerliğine tehdit oluşturmakta ve ölçmede sistematik bir hataya neden olmaktadır (Clauser ve Mazor, 1998; Osterlind, 1983). Angoff'un maddelerin belirli bir gruba görece zor gelmesi (madde güçlük indekslerinin görece düşük olması) durumunda madde yanlılığından söz edilebileceği düşüncesine benzer şekilde, Scheuneman (1979) çalışmasında madde yansızlığını ölçülen özellik bakımından tüm bireylerin bir maddedeki doğru yanıt oranının, onların ait olduğu (etnik) gruptan bağımsız olarak, aynı olması şeklinde tanımlamıştır. Kelderman (1989) madde yanlılığını, aynı yetenek düzeyinde ancak farklı gruplarda yer alan bireylerin doğru yanıt verme olasılıklarının birbirinden farklı olması şeklinde tanımlamaktadır. Pine (1977), bir test maddesinin yansız kabul edilebilmesi için bireyin grup üyeliğinden bağımsız olarak ölçülmek istenen tek boyutlu özelliğe sahip tüm bireylerin maddeyi doğru yanıtlama olasılıklarının eşit olması gerektiğini savunmaktadır.

Madde yanlılığı ve bireylerin bir madde üzerindeki değişen performansı konularının tarihçesi, standart testlerin başlangıcına kadar gitmektedir (Bezruczko vd., 1989). Örneğin, Binet ve Simon (1905) zihinsel yetkinlik üzerine yürüttükleri çalışmalarının sonucunda işçi sınıfı ailelerin çocukları ile yüksek sosyal statüye sahip ailelerden gelen çocukların testten elde ettikleri puanların farklılaştığını raporlamıştır. Bunu destekleyecek biçimde zihinsel yetkinlik ya da zekâ testlerinde özellikle bazı ırkların aleyhine sonuçların elde edildiği (Jensen, 1973; Jensen, 1976) ve bu konuda yorumlamalar yapılırken dikkatli olunması gerektiği bilinmektedir (Scarr ve Weinberg, 1976). Gould (1981) da özellikle zekâ testlerinin ırkçı bir anlayışla kullanılmasına yönelik eleştirilerini dile getirmiştir. Bu noktada, ölçülmek istenen psikolojik yapının doğru bir şekilde ölçülüp ölçülmediği konusunda emin olunamadığı durumlarda, testlerden elde edilen puanlar üzerinden bireyler/gruplar arasında karşılaştırma yapmak uygun değildir.

Testler ve maddeler üzerinde yanlılık çalışmalarına yönelik pratik uygulamaların 1960'lı yılların sonlarına doğru hızlandığı ve yaygınlaşmaya başladığı görülmektedir. Bu kapsamdaki çalışmaların sayıları, psikolojide, sosyal bilimlerde ve eğitim bilimlerindeki etkilerinin fark edilmesiyle giderek artmaya başlamıştır (Gómez-Benito vd., 2018). Bir testte yer alan yanlı olma potansiyeli taşıyan maddeleri belirlemek için günümüzde yaygın olarak kullanılan istatistiksel tekniklerin birçoğunun 1970'lerde geliştirilmeye başlandığı görülmektedir. Penfield ve Camilli (2007), bu istatistiksel teknikleri, tarihsel geri planlarını da dikkate alarak açıklamaktadır. Holland ve Thayer'in (1988) kapsamlı ve derleyici çalışmaları gibi birçok değerli katkı sayesinde, yanlı olma potansiyeli taşıyan maddeleri ve değişen madde fonksiyonu (DMF) belirleme çalışmalarında ciddi mesafeler katedildiği görülmektedir. DMF, bir testi alan gruplar arasındaki farklı yanıt örüntülerini belirlemek için kullanılan bir istatistiksel tekniktir ve potansiyel olarak yanlı test maddelerinin belirlenmesinde kullanılmaktadır. Bir maddenin gruplar arasında yanlılık içerdiği kararının verilebilmesi için öncelikle belirgin bir DMF gözlenmiş olması gerekmektedir. Bununla birlikte DMF, madde yanlılığı kararının verilebilmesinde yeterli bir gösterge değildir. Gözlenen DMF'nin madde yanlılığı kaynaklı olduğunun belirlenmesine yönelik devam çalışmalara ihtiyaç vardır (Zumbo, 1999). DMF gözlenmesi ve bir maddenin yanlılık içerdiği kararının verilmesi açıkça ayrı iki durumdur.

DMF çalışmaları sıklıkla, yeni bir ölçme aracı geliştirirken, ölçme aracını farklı kültürlere uyarlarken ya da daha genel olarak test puanlarına bir geçerlik kanıtı sunmak adına yürütülmektedir (Zumbo, 2007). DMF çalışmalarına yaygın olarak eğitim bilimleri alanında rastlanılmaktadır (Li ve Zumbo, 2009). Buna karşın, psikoloji (Dodeen ve Johanson, 2003), sağlık bilimleri (Gelin vd., 2004; Iwata vd., 2002) ve spor bilimleri (Gao ve Zhu, 2009) gibi farklı alanlarda da DMF analizleri kullanılabilmektedir.

Özetle testte yanlı maddelerin bulunması testten alınan puanlarda geçerlik sorunlarına yol açabilmektedir. Bu nedenle DMF belirleme çalışmaları geçerlik bağlamında değerlendirilmektedir. Bir testteki bir maddenin ya da maddelerin DMF'li olarak belirlenmesinin ardından en az üç konunun tartışmaya açık hale geleceği ifade edilmektedir (de Ayala, 2009):

1) Bir madde ya da maddeler istatistiksel hesaplamalar sonucunda DMF'li olarak işaretlendi ise panel çalışmaları ile uzmanlardan görüş alınmalıdır. Maddenin yanlı olduğunun söylenebilmesi adına bilimsel kanıtlar sunulmalıdır. DMF sonucu tek başına maddenin yanlı olduğu yorumunun yapılması için yeterli değildir.

2) Tespit edilen DMF'nin miktarı ve gücünün ne olduğu önemlidir. Gruplar arasındaki farklılığın ihmal edilebilir bir düzeyde mi yoksa ciddi miktarda mı olduğu tartışılmalıdır.

3) DMF'li maddelerin analizlere etkisinin ne olduğu tartışılmalıdır.

Bu konular alan yazınında hala tartışılmaktadır. Ek olarak, DMF incelemeleri çeşitli şekillerde gerçekleştirilebilir ve farklı yöntemlerin aynı veri setinin analizinde işe koşulması durumunda farklı sonuçlar elde edilebilir (farklı sayıda maddede DMF gözlenmesi, farklı maddelerin DMF'li olarak işaretlenmesi gibi). Bu durum DMF ve madde yanlılığı belirlemede kullanılacak yöntem ya da yöntemlerin belirlenmesinde dikkatli olunması gerektiğini göstermektedir (Ironson ve Subkoviak, 1979). Bu kapsamda ilgili alan yazınında DMF belirlemede kullanılabilecek yöntemlerin farklı koşullarda performanslarının karşılaştırıldığı, hangi yöntemin hangi durumlarda daha uygun olduğu üzerine birçok araştırma görmek mümkündür (Raju vd., 1993; Wainer vd., 2010). Bir ölçme aracındaki DMF'nin incelenmesinde tek bir DMF belirleme yönteminin kullanıldığı çalışmalar (Hauger ve Sireci, 2008; Ozarkan vd., 2017) olduğu gibi birden fazla sayıda yöntem ile DMF'nin incelendiği çalışmalar (Dzul-Garcia ve Atar, 2020; Tat ve Doğan, 2018) da mevcuttur. Ayrıca, analiz için seçilen DMF yönteminden bağımsız olarak DMF'li maddelerin tespitinin de ötesinde DMF'nin kaynaklarının incelenmesi beklenmektedir (Anastasi ve Urbina, 1997). Kısaca, DMF analizleri testte yanlı davranan bir madde bulunup bulunmadığını incelemek için yürütülmektedir ve DMF'nin tespiti için en az bir DMF belirleme yönteminin işe koşulması gerekmektedir.

1.2. DMF Belirleme Yöntemleri

DMF belirlemek için geliştirilen ilk istatistiksel yöntemlerde madde güçlüklerine (Angoff'un DMG Yöntemi gibi), kontenjans tablolarına (Mantel-Haenszel Yöntemi gibi) ve gruplar arası varyansa odaklanıldığı görülmektedir (Camilli ve Shepard, 1994). Özellikle ikili puanlanan maddeler üzerinden iki grubun karşılaştırılması için kullanılan DMF belirleme yöntemlerine DMG (Delta Plot Yöntemi; Angoff ve Ford, 1973), Standardizasyon (Dorans'ın Standardizasyonu; Dorans ve Kulick, 1986), Mantel-Haenszel Yöntemi (M-H; Holland ve Thayer, 1988), Lojistik Regresyon (Swaminathan ve Rogers, 1990) ve SIBTEST (Shealy ve Stout, 1993) örnek verilebilir. Bunlar haricinde yapısal eşitlik modellemesi (YEM) yaklaşımına dayalı MIMIC (Multiple Indicators Multiple Causes; Finch, 2005) ve rastgele etkiler (random-effects) modelleri (Gamerman vd., 2018) ile de DMF incelemelerinin yapılması mümkündür. Bu yöntemler haricinde Madde Tepki Kuramına (MTK) dayalı yöntemler de geliştirilmiştir. Lord'un χ^2'si (Lord, 1980), MTK Olabilirlik Oran (MTK-OO; Thissen vd., 1988, 1993), Raju'nun İşaretli ve İşaretsiz Alan İndeksleri (Raju, 1988, 1990) ve Lordif (Logistic ordinal regression differential item functioning; Choi vd., 2011) bu yöntemlere örnek verilebilir. İlerleyen süreçte bu yöntemlerin bazı sınırlılıklarını aşmak için genelleştirilmiş veya yeni yöntemler de geliştirilmeye devam edilmiştir.

Bilindiği üzere DMF belirleme çalışmalarında kullanılmak üzere çok sayıda yöntem önerilmiştir. Çok sayıda yöntem bulunmasının temel nedeni, yöntemlerin kendi içlerinde güçlü ve sınırlı yanlarının bulunmasıdır. Örneğin, Angoff'un DMG yöntemi hesaplama kolaylığı ile öne çıkarken, SIBTEST oldukça karmaşık bir algoritmaya sahiptir. Geleneksel yöntemlerde görece daha küçük örneklemler ile çalışılırken, MTK'ya dayalı yöntemler ile çalışılacak ise analizler için daha büyük örneklemler gerekmektedir. DMF belirleme yöntemlerinin büyük çoğunluğu tek bir uygulamadan elde edilen veriler ile DMF analizi gerçekleştirmeye olanak sağlarken rastgele etkiler (random-effects) modeli için en az iki uygulamanın gerçekleştirilmesi gerekmektedir. Bu çalışmada DMF belirleme yöntemlerinden Angoff'un DMG yönteminin özellikleri açıklanmıştır.

1.3. Angoff'un Dönüştürülmüş Madde Güçlükleri Yöntemi

Angoff'un Dönüştürülmüş Madde Güçlükleri (DMG) yöntemi erken dönem örneklerinden biri olarak öncü bir role sahiptir. Daha sonraları çok sayıda DMF belirleme yöntemi geliştirilmiş ve bu yöntemler çeşitli biçimlerde sınıflandırılmıştır (Örneğin; Elosua ve Wells, 2013; Wainer, 1993). Bu perspektiften bakıldığında, Angoff'un DMG yöntemi gözlenen puana ve Klasik Test Kuramı'na (KTK) dayalı ve ampirik tabanlı bir yöntem olarak değerlendirilebilmektedir. Angoff'un DMG yöntemi DMF belirleme alan yazınında kısa süre içinde popülerlik kazanmıştır. Angoff'un daha önceleri de yanlılık üzerine çalışmaları olduğu bilinmektedir. Bununla birlikte 1972 yılında kültürel farklılıkları inceleme çalışmalarında kullanılabilecek bir yöntem üzerine önerileri dikkat çekmiştir. Ayrıca, Angoff'un yönteminin temelleri Thurstone'un (1925) mutlak ölçekleme (absolute scaling) yöntemi ile de ilişkilendirilebilmektedir. Angoff'un yöntemi ise özellikle kültürler arası çalışmalardaki kullanılabilirliği açısından dikkat çekmiştir ve Angoff geliştirdiği yöntemi farklı çalışmalarında (Angoff, 1975; Angoff ve Cook, 1988; Angoff ve Modu, 1973) kullanmıştır.

Dönüştürülmüş Madde Güçlükleri yönteminde basitçe her madde için farklı gruplardaki (örn. cinsiyet) bireylerin verdiği doğru yanıtların oranları karşılaştırılmaktadır. 1-0 şeklinde ikili puanlanan (dichotomous) maddelerde doğru yanıtlama oranları madde güçlüklerine (item difficulty) eşittir. Angoff'un DMG yönteminde basit bir hesaplama ile elde edilen çıktılar bir grafik aracılığıyla görselleştirilebilmekte ve kolay bir şekilde yorumlanabilmektedir (Angoff, 1982; Osterlind, 1983). Bu grafiğe Delta Grafiği ismi verilmektedir. Bu nedenle yöntem, Angoff'un Delta Grafiği / Delta Plot yöntemi (Shepard vd., 1985) şeklinde de isimlendirilebilmektedir.

Yöntemin daha rahat anlaşılabilmesi için örnek bir durum ele alalım. Matematik başarısını ölçen bir testin uygulanmasının ardından, elimizde öğrencilerin yanıtları ve cinsiyetlerine ait bilgiler bulunsun. Testte yer alan maddelerden bir tanesinin yanlı olabileceği şüphesi ile DMF analizi gerçekleştirilecek olsun. Bu durumda öğrencilerin maddelerden aldığı puanlardan (0 ve 1) oluşan matrisin yanında cinsiyetlerini belirten 0 ve 1 etiketlerinden oluşan bir veri seti oluşturalım. 1 kadınları, 0 erkekleri temsil etmek üzere 1 grubundan (kadınların) elde edilen ilgili maddenin madde güçlüğü (p) değeri ile 0 grubundan (erkek) elde edilen madde güçlüğü değeri karşılaştırılacaktır. Aradaki farkın belli bir değerden fazla olması durumunda, madde DMF'li olarak işaretlenir. Aslında doğrudan iki gruptaki madde güçlüğü değerleri karşılaştırılmamakta, basit bir doğrusal dönüşüm kullanılarak p değerlerinden Δ değerleri elde edilmektedir. Ama algoritmanın mantığı basitçe bu karşılaştırmaya dayanmaktadır. Tabi bu işlemler için Angoff'un önerdiği bir algoritma mevcuttur. Algoritma, bu çalışmanın yöntem bölümünde ayrıntılarıyla açıklanmıştır.

Angoff'un DMG yönteminin alan yazınındaki popülaritesini kaybetmesinin nedeni yalnızca yeni DMF belirleme yöntemlerinin alana kazandırılmış olması değil, aynı zamanda yöntemin ciddi sınırlılıklarının bulunduğunun fark edilmesidir. Bu çalışmanın amacı DMF belirleme çalışmalarında çeşitli yönlerden sık sık eleştirilen, buna karşın günümüzde hala bazı çalışmalarda kullanılan Angoff'un DMG yönteminin algoritmasını yakından inceleyerek, çalışmalarda kullanılabilirliğinin sorgulanması ve eleştirilerin nedenlerini irdelemektir. Bu amaçla aşağıdaki sorulara yanıt aranmıştır.

1) DMF belirlemede Angoff'un DMG yönteminin güçlü ve sınırlı yanları nelerdir?

2) Yöntemin hangi çalışmalarda/alanlarda, hangi sınırlılıklarla kullanılması daha uygundur?

2. YÖNTEM

2.1. Araştırma Deseni

Bu araştırmada var olan bir durumun var olduğu şekliyle ortaya konulması amaçlandığından temel araştırma (basic research) olarak nitelendirilebilir (Fraenkel vd., 2012).

2.2. Angoff'un DMG Yönteminin Algoritması

Angoff'un DMG yöntemine göre DMF gösteren bir maddenin belirlenmesi için yapılacak hesaplamalar elle ya da temel bir elektronik tablo programı ile gerçekleştirilebilir. Bu bölümde Angoff'un DMG yönteminin algoritması basamak basamak açıklanacaktır.

1. Öncelikle karşılaştırılacak gruplar belirlenmelidir. Bu gruplardan biri odak, diğeri ise referans grubu olacaktır.[4] Örneğin, bir test uygulamasının ardından testin maddelerini yanıtlayan bireylerin yanıtlarının yanında cinsiyet bilgilerinin de alındığını düşünelim. Odak grubu kadınlar, referans grubu erkekler olarak seçildiği durumda, kadın bireylere 1, erkek bireylere 0 ataması yapılacaktır.

2. Her bir madde için her bir grupta (odak ve referans) madde güçlük indeksi ayrı ayrı hesaplanmalıdır. Bu hesaplama oldukça kolaydır. Madde güçlük indeksi (p_j) maddelerin 1 ve 0 şeklinde ikili puanlandığı (dichotomous) durumda maddeden alınan 1 puanlarının sayısının, toplam katılımcı sayısına bölümü ile hesaplanmaktadır. Bu yöntemde gruplar söz konusu olduğu için gruplar g alt indisi ile gösterilmek üzere madde güçlük indeksleri p_{jg} ile gösterilmiştir.

3. Madde güçlük indeksleri, normalleştirilmiş z puanlarına (kısaca z puanları) dönüştürülmelidir. Bu aşamada, elde edilen p_j değerleri, z puanları tablosunda karşılık gelen değerlere dönüştürülür. Örneğin,

[4] Odak (focal) grup ve referans (reference) grup tanımlaması günümüzde oldukça yaygın bir şekilde kullanılmasına karşın, bazı çalışmalarda bu gruplar çalışma (study) grubu – temel (base) grup (Dorans ve Kulick, 1983), odak grup – temel grup (Dorans, 1989), azınlık (minority) – çoğunluk (majority) grupları (de Ayala, 2009) şeklinde de ele alınabilmektedir. Türkçede odak grubunun çevirisinin fokal grup olarak yapıldığı bir çalışma da mevcuttur (Korkmaz, 2006).

0,95 ve 0,975 değerleri sırasıyla -1,64 ve -1,96 değerlerine karşılık gelecektir. Burada farklı gruplarda p_{jg} değerlerinin dönüştürülmesi ile elde edilen puanlar z_{jg} ile gösterilmiştir.

4. Elde edilen z_{jg} puanları Δ değerlerine dönüştürülmelidir. Bu aşamada, Angoff'un önerdiği doğrusal bir dönüşüm söz konusudur. Buna göre, her bir gruptaki her bir z puanı, $\Delta_{jg} = 4z_{jg} + 13$ eşitliği ile bir Δ değerine dönüştürülür (ortalama 13, standart sapma 4). Bu dönüşüm ile negatif değerler yerine pozitif değerler ile işlem yapma kolaylığı elde edilmiş olur. Örneğin, sıklıkla -3 ile +3 arasında değişen z değerleri Δ değerlerine dönüştürüldüğünde 1 ile 25 arasında değerler elde edilir.

5. İki ayrı grup için elde edilen Δ değerleri $(\Delta_{j0}, \Delta_{j1})$ bir saçılım grafiği (scatter plot) aracılığı ile incelenir. İki ayrı grup için elde edilen delta değerleri, referans grubun yatayda, odak grubun dikeyde yer aldığı bir dağılım grafiği ile incelenebilir. Bu grafiğe özel olarak delta plot adı verilmektedir. Grafikte yer alan her bir nokta delta noktası (delta point) olarak isimlendirilir. Bu noktadaki grafiksel incelemede, eğer testte DMF'li madde yer almıyorsa noktaların birbirine yakın olması ve elipse benzer bir şekil oluşturması beklenir. Eğer noktalar birbirinden uzaklaşıyor ve elipsoit şekil bozuluyorsa, testte karşılaştırılan gruplar açısından DMF'li madde bulunması olasıdır. Çünkü elipsoitin bozulması, madde × grup etkileşiminin varlığı şeklinde yorumlanmaktadır ve bu durumda testteki bir madde diğer maddelere göre görece, bir gruba kolay gelirken diğerine zor gelmiştir (Devine ve Raju, 1982). Angoff (1975) grafikte sapan noktaların farklı gruplardaki bireyler için farklı psikolojik anlamlar ifade edebileceğini söylemektedir.

6. Dik uzaklıklar hesaplanmalıdır. Grafiksel incelemenin yanı sıra, maddelerin DMF gösterip göstermediğinin belirlenmesi için delta noktalarının ana eksenine (major axis) olan dik uzaklıklarının (perpendicular distance) hesaplanması gereklidir. Elipsin ana ekseni $\Delta_{j1} = a + b\Delta_{j0}$ denklemi ile ifade edilir. Bu noktada kesişim (a) ve eğim (b) parametrelerinin hesaplanması gerekmektedir.

$b = \frac{s_1^2 - s_0^2 + \sqrt{(s_1^2 - s_0^2)^2 + 4s_{01}^2}}{2s_{01}}$ ve $a = \bar{x}_1 - b\bar{x}_0$ eşitlikleri yardımıyla gerçekleştirilen hesaplamalar[5] sonucunda kesişim ve eğim parametreleri hesaplanır. Burada, $\bar{x}_1$ ve $\bar{x}_0$ sırasıyla odak ve referans gruptaki ortalamaları, s_1^2 ve s_0^2 sırasıyla odak ve referans gruptaki varyansları ve s_{01} kovaryansı göstermektedir.

Delta noktalarının ana eksene uzaklıklarını ifade eden dik uzaklıklar (D_j) ise aşağıdaki şekilde hesaplanır:

$$D_j = \frac{b\Delta_{j0} + a - \Delta_{j1}}{\sqrt{b^2 + 1}}$$

Eşitlikten de anlaşılacağı üzere, noktaların bir doğruya olan uzaklıkları hesaplanmaktadır. Büyük uzaklık değerleri, delta noktalarının ana eksenden uzaklaştığını göstermektedir ve bu durumda maddenin DMF göstermesi söz konusu olabilir. Uzaklık değerleri negatif ya da pozitif olabilir. Pozitif büyük değerler, ilgilenilen maddenin referans gruptaki bireylere odak gruptaki bireylere göre daha kolay geldiğini gösterirken, negatif büyük değerler (mutlak değer açısından), ilgili maddenin odak gruptaki bireylere referans gruptaki bireylere göre daha kolay olduğunu gösterir.

Dik uzaklıkların DMF şüphesi oluşturacağı durumun ne olması gerektiği konusunda çeşitli öneriler bulunmaktadır. Bu öneriler teknik açıdan sınıflandırma amacı taşımaktadır (madde DMF'li ya da değil). Önerilen eşik değerler istatistiksel bir manidarlık düzeyine atıfta bulunmamaktadır, bunun yerine bir etki büyüklüğü olarak değerlendirilebilirler. Alan yazınında sıklıkla, Mantel-Haenszel yönteminin yorumlanmasında kullanılan, C düzeyi DMF'yi işaret eden 1,5 eşik değeri (Zwick ve Ercikan, 1989) Angoff'un DMG yönteminde de kullanılmaktadır. Holland ve Wainer (1993) 1,5 eşik değerinin DMF

[5] Magis ve Facon'un (2012) çalışmasında bu formül hatalı olarak verilmiştir. Bu hata, Magis ve Facon'un bir başka çalışmasında (2014) düzeltilmiştir.

incelemelerinde sıklıkla kullanıldığını ifade etmektedir ve araştırmacılar da [-1,5; 1,5] aralığının dışında kalan uzaklıkları, çalışmalarında DMF olarak değerlendirmektedir (Facon ve Nuchadee, 2010; Robin vd., 2003).

2.3. Örnek Uygulama İçin Veri Seri Üretimi

Gerçek veri seti üzerinde gerçekleştirilecek analizler sonucu elde edilen bulguların hatalı şekilde yorumlanmasına neden olmamak için simülasyon verisi ile çalışılması uygun görülmüştür. Bu amaca yönelik olarak R'daki 'psych' paketi (Revelle, 2023) yardımıyla bir veri seti üretilmiş ve DMF'li maddenin daha rahat görülebilmesi için üzerinde bazı düzenlemeler yapılmıştır. Veri setinin oluşturulmasında kullanılan satır komutları Ek-1'de verilmiştir.

3. BULGULAR

Bu bölümde, yöntem bölümünde detaylı bir şekilde açıklaması verilen algoritma ve alan yazınındaki bilgiler üzerinden sırasıyla Angoff'un DMG yönteminin güçlü ve sınırlı yanları sunulmuş, ardından hangi araştırmalarda, hangi şekilde kullanılabileceğine ilişkin bilgi verilmiştir.

3.1. Angoff'un DMG Yönteminin Güçlü Yanları

Angoff'un DMG yöntemi, anlaşılması oldukça kolay bir yöntemdir. Bu yöntemde temel olarak bir maddenin bir gruptaki bireylere diğer gruptaki bireylerden daha kolay gelmesi (madde güçlük indeksinin düşük değer alması) maddenin yanlı olabileceğine ilişkin bir öngörü sağlamaktadır. Yöntemin algoritmasının basitliği (Oosterhof vd., 1984) onun güçlü yanlarından biridir.

DMF belirleme çalışmalarında örneklem büyüklüğü maddelere ilişkin istatistikleri etkileme gücüne sahiptir. Özellikle bazı yöntemlerde (MTK'ya dayalı DMF belirleme yöntemleri gibi) görece büyük örneklemlerle çalışılması bir zorunluluk olarak görülmektedir. Buna karşın Angoff'un DMG yönteminde küçük örneklemler ile de çalışmalar yürütülebilir (Muñiz vd., 2001). Örneğin, sınıf içinde uygulanan bir ölçme aracı üzerinde DMF analizleri yürütüleceği zaman, küçük bir örneklem ile çalışılması gerekmektedir ve bu durum da seçilebilecek DMF belirleme yöntemine kısıt getirecektir.

Yöntemin içerdiği matematiksel işlemler oldukça sadedir. Algoritmada hesaplanan delta değerlerinin, ana eksen ve dik uzaklıklarının hesaplanması kolaydır ve delta değerleri (noktaları) bir grafik ile görselleştirilebilir (Angoff, 1982). Grafiksel incelemenin sonucunda yorumlamalar kolaylıkla yapılabilir. Bu nedenlerle, Angoff'un DMG yöntemi, pek çok sınırlılığına rağmen hala araştırmalarda kullanılabilmektedir.

3.2. Angoff'un DMG Yönteminin Sınırlı Yanları

Angoff'un Dönüştürülmüş Madde Güçlükleri yöntemi, gruplar arasında bir maddeyi doğru yanıtlama oranlarına odaklanmaktadır. Bu yöntemde madde ayırt edicilikleri (item discrimination) göz önünde bulundurulmaz. Eğer, bir madde gruplar arasında madde ayırt ediciliği açısından farklılaşıyorsa (tek biçimli olmayan DMF) bu yöntem DMF'yi tespit etmek için uygun bir yöntem olmayacaktır. Bu yöntemin önemli sınırlılıklarından biri budur.

Bu yöntemde bir maddenin bir gruptaki herkes tarafından doğru ya da herkes tarafından yanlış yanıtlanması matematiksel açıdan bir soruna neden olabilir. Çünkü bu durumda p_j değeri 1 ya da 0 olacaktır. Bu durumda delta değerleri sonsuz olacaktır. Angoff ve Ford (1973) bu sınırlılığı aşmak için madde güçlüklerini [,05; ,95] arasına sıkıştırmayı önermiştir. Nitekim, bu yöntemin en önemli sınırlılığı bu değildir.

Yöntemin önemli sınırlılıklarından biri madde etkisinin (item impact) varlığı ve madde ayırt ediciliğidir. Hunter (1975), Lord (1977) ve Rudner'in (1978) ifade ettiği üzere, cevaplayıcılar arasında ölçülen özellik açısından farklılık söz konusu olabilir ve bu durumda gruplar farklı ortalama yeterlik düzeyine sahip oldukları için ortalama madde güçlükleri farklılaşır. Bu halde, madde etkisi ortaya çıkmasına rağmen, yöntem maddeyi DMF'li olarak işaretleyecektir. Bu durum, sonraları Angoff (1993) tarafından da kabul

edilmiştir. Gruplardaki bireylerin yeterlik düzeylerinin dikkate alınmaması durumunda oluşan madde güçlükleri farkı, tıpkı t-testindeki ya da ANOVA'daki gibi gruplar arasındaki farktan kaynaklanmaktadır, maddenin yanlılığından değil. Basitçe ifade etmek gerekirse, *p* değerleri (aynı zamanda delta değerleri gibi düşünülebilir) madde güçlüklerini yansıttığı gibi grup farklılıklarını da yansıtmaktadır. Bu durum, yöntemde madde ayırt ediciliklerinin göz önüne alınmamasından kaynaklanmaktadır. Yöntemin, yalnızca madde güçlüklerine odaklanıyor oluşuna gelen eleştiriler neticesinde Angoff (1982) hesaplamalara madde ayırt ediciliklerini de dahil edecek şekilde bir düzenleme yapmıştır. Revize Edilmiş Dönüştürülmüş Madde Güçlükleri Yöntemi (Revised TID Method) bu şekilde oluşmuştur. Ancak bu yöntem de madde yanlılığı incelemeleri için yetersiz görülmüştür (Seong ve Subkoviak, 1987).

Angoff'un DMG yönteminde eşik değerler (threshold) ile ilgili bir sınırlılık da söz konusudur. Bu yöntemde bir maddenin DMF'li olup olmadığına maddeye ait delta değerinin ana eksene dik uzaklık değerine göre karar verilmektedir. Bu uzaklığın ne kadar olduğu durumda maddenin DMF'li olarak işaretleneceği ise bir tartışma konusudur. Alan yazınında sıklıkla 1,5 değeri sınır değer olarak kullanılmaktadır, buna karşın Angoff net bir eşik değeri önermemiştir. Oysa, eşik değerin kaç olacağı önemli bir konudur. Yüksek bir eşik değer seçildiğinde DMF'li maddeler tespit edilemeyebilir, düşük bir değer seçildiğinde ise DMF'li olarak işaretlenen maddelerin yanlı olduğuna ilişkin öne sürülen kanıtlar bilimsel bir zemine oturtulamayabilir.

Bir ölçme aracındaki maddeler 1-0 şeklinde puanlanabileceği gibi iki kategoriden daha fazla sayıda puan kategorisi de mevcut olabilir. Angoff'un DMG yöntemi yalnızca ikili puanlanan (dichotomous) maddeler için kullanılabilir, çoklu puanlanan (polytomous) maddeler için bu analiz gerçekleştirilememektedir. Ayrıca, bu yöntemde yalnızca iki grup karşılaştırılabilmektedir.

DMF iki türlü olabilir: tek biçimli (uniform) ve tek biçimli olmayan (non-uniform). Bir gruptaki bireylerin bir maddeye doğru yanıt verme olasılığı, tüm yetenek düzeylerinde diğer gruptaki bireylerin olasılıklarından yüksek ise bu tek biçimli DMF'ye bir örnektir. Buna karşın, bir gruptaki bireylerin bir maddeye doğru yanıt verme olasılığı bazı yetenek düzeylerinde diğer gruptaki bireylerin olasılıklarından düşük, bazı yetenek düzeylerinde yüksek ise bu durum tek biçimli olmayan DMF'ye bir örnek oluşturacaktır (Swaminathan ve Rogers, 1990). Angoff'un DMG yönteminde yetenek grupları arasında bir eşleme yapılmadığı için, bu yöntem ile yalnızca tek biçimli DMF'nin tespit edilmesi hedeflenmektedir.

Angoff'un DMG yönteminin pek çok sınırlılığı bulunmasına karşın, günümüzde hala kullanıldığını söylemek mümkündür. Örneğin, Aituariagbon ve Osarumwense (2022) tarafından yürütülen bir çalışmada Mantel-Haenszel, Standardizasyon ve Angoff'un DMG yöntemleri ile SSCE 2019 Ekonomi isimli bir testte DMF incelemesi yapılmıştır. Yazarlar, Standardizasyon ve Angoff'un DMG yöntemlerinin DMF belirlemede daha uygun yöntemler olabileceğini raporlamışlardır. de Ruiter ve Bers (2022) tarafından bir ölçme aracı geliştirme sürecinde geçerliğe ilişkin kanıtlar sunmak adına cinsiyet ve yaş grupları arasında DMF analizleri yürütülmüştür ve bu aşamada ilk olarak Angoff'un DMG yöntemi kullanılmıştır. Araştırmacılar yöntemin sınırlı yönlerinin farkında olarak ardından M-H yöntemi ile de DMF incelemesi yapmışlardır. Yöntem bu çalışmada, grafiksel inceleme imkânı sunduğu için DMF hakkında bir öngörü vermesi amacıyla kullanılmıştır. Farcomeni ve arkadaşları (2022) tarafından bir ölçeğin Avrupalı örnekleminde kullanımına yönelik bir çalışma yürütülmüştür. Bu çalışmada ölçeğin bir ölçme kuramına uygunluğunun yanı sıra DMF incelemesi de yapılmıştır. Bu aşamada DMF belirleme yöntemi olarak Angoff'un DMG yönteminden yararlanılmıştır. Van Vo ve Csapó (2023) tarafından yürütülen bir çalışmada bir testin farklı uygulama koşullarındaki sonuçları üzerinden elde edilen ölçümlerin geçerliğine bir kanıt sunmak adına Angoff'un Delta Plot (DMG) yönteminden yararlanılmıştır.

3.3. Hangi Araştırmalarda Kullanılabilir?

Bu çalışmanın yöntem bölümünde sunulduğu üzere, Angoff'un DMG yöntemi, bir testteki maddelerin farklı gruplardaki bireylere madde güçlüğü açısından avantaj sağlayıp sağlamadığını test etmektedir. Temel bir kavrayışla, testteki maddeler bir gruptaki bireyler tarafından diğer gruptaki bireylere göre daha kolay yanıtlanabiliyorsa, maddelerin DMF'li olabileceği konusunda bir öngörü sağlamaktadır. Ancak, bu yaklaşım günümüzdeki DMF analizi bakış açısından kısmen sınırlıdır. Çünkü, maddelerin güçlük parametrelerinin dışında da parametreleri bulunmaktadır. Bunlar göz önünde bulundurulmadığı takdirde, hatalı yorumlamalar söz konusu olabilir. Ayrıca, ölçülen özellik bakımından bir gruptaki bireyler diğer gruptaki bireylere göre halihazırda daha iyi durumda ise, bu noktada maddelerin güçlük parametreleri gruplar arasında farklılaşacaktır ve maddeler DMF'li olarak işaretlenecektir. Oysa, bu durum gruplar arasındaki gerçek farktan kaynaklanmaktadır. Yöntemin yalnızca madde güçlüklerine odaklanması teorik sınırlılığından dolayı, bu yöntem ancak karşılaştırılan iki grubun gerçek ortalamalarının birbirine eşit ya da çok yakın olduğu durumlarda daha işlevsel sonuçlar verebilir (Shepard vd., 1981). Bu yönteme göre ayrıca, bazı maddelerin diğer maddelere göre daha ayırt edici olması durumunda da yöntem maddeleri DMF'li olarak işaretleyecektir.

Angoff'un DMG yönteminden sonra geliştirilen DMF belirleme yöntemlerinde gruplar arasında karşılaştırma yapılırken yetenek eşlemesi yapıldığı görülmektedir. Çünkü, grup ortalaması üzerinden karşılaştırma yapmak, gruplardaki alt ve üst başarı gruplarındaki bireylerin hatalı karşılaştırılmasına neden olabilir. Angoff'un DMG yöntemi ile DMF analizinde yetenek eşlemesi yapılmadığı ve karşılaştırmaların grup ortalamaları üzerinden yapılması nedeniyle bir sınırlılığı bulunmaktadır.

Özetle, Angoff'un DMG yöntemi ile DMF analizi, önemli araştırmalarda tek başına kullanılmamalıdır. Ancak, küçük örneklemlerle çalışılırken ve gruplar arasında ölçülen özellik açısından ciddi bir farklılaşma beklenmediği durumlarda bir öngörü vermesi amacıyla sınırlı bir kullanımı düşünülebilir. Ek olarak, test/madde yanlılığı ve DMF gibi konularda, özellikle DMF analizinin bireylere açıklanmasında kullanılabilecek en iyi DMF belirleme yöntemlerinden biri olduğu söylenebilir.

3.4. Örnek Bir R Uygulaması

Bu çalışmanın temel amaçlarından biri Angoff'un DMG yöntemi ile DMF belirlemenin nasıl gerçekleştirildiğinin aktarılmasıdır. Bu nedenle, örnek bir veri seti üzerinden analiz gerçekleştirilmiş ve analiz basamakları adım adım açıklanmıştır. Bu analiz, elle ya da bir elektronik tablo programı ile yapılabileceği gibi, sosyal bilimler alanında yaygın bir şekilde kullanılan R yazılımı (R Development Core Team, 2023) ile de gerçekleştirilebilmektedir. DMF analizleri için daha önceden geliştirilmiş pek çok yazılım bilimsel araştırmalarda yaygın olarak kullanılmıştır. Bu yazılımlarda bu çalışmada açıklanan Angoff'un DMG yöntemi ile DMF analizi gerçekleştirilemediği gibi, yazılımlar günümüzdeki istatistiksel analiz anlayışından kısmen uzaktır. Bu nedenle, bu çalışmada kullanıcıların dilediklerinde kendi satır komutlarını girerek ya da diğer geliştiriciler tarafından geliştirilen paketleri kullanarak kolaylıkla analizler yürütebilecekleri, özgür bir yazılım olan R tercih edilmiştir. Angoff'un yöntemi ile analizi örneklendirmek için R'da bir veri seti üretilmiştir ve bu veri setinin üretimi ile ilgili bilgiler çalışmanın ilgili bölümünde verilmiştir. Bu bölümde analizlerin satır komutları ile adım adım nasıl yapılacağı gösterilmiştir.

1. R ortamında ön düzenlemeler. Veri seti analiz için hazır hale getirildikten sonra (sütunlarda maddelerden alınan puanlar ve grup değişkeni yer alacaktır) R'da yeni bir betik açılarak ilgili paketlerin indirilmesi ve çalıştırılması gerekmektedir. R'da pek çok temel paket ve fonksiyon bulunduğu gibi, başka geliştiriciler tarafından geliştirilen paket ve fonksiyonların kullanılması da mümkündür. Bu çalışmada Angoff'un DMG yöntemini içeren kapsamlı bir paket olan '`difR`' paketi (Magis vd., 2010) kullanılmıştır. Paketin indirilmesi ve R ortamında kullanıma hazır hale getirilmesi için, R'ın komut penceresine aşağıda verilen satır komutları girilmelidir.

```
install.packages("difR")
library(difR)
```

Yukarıda verilen satır komutları girilip çalıştırıldıktan sonra ilk olarak '`difR`' paketi indirilmiş, ardından aktif hale getirilmiş olacaktır. Bu çalışmada kullanılan veri setinde (data) 200 satır ve 21 sütun bulunmaktadır. İlk 20 sütunda maddelerden alınan puanlar, 21. sütunda ise cinsiyet bilgileri (1 ve 0) yer almaktadır. R ortamına yükleme işlemi başarılı bir şekilde gerçekleştirildiğinde `dim(data)`, `head(data)`, `str(data)` gibi komutlarla, verinin yapısı görülebilir.

2. Angoff'un DMG yöntemi ile analizin gerçekleştirilmesi. Kullanılan ilgili paketteki tek bir fonksiyon yardımı ile analiz gerçekleştirilebilmektedir. Bunun için R Düzenleyici penceresine aşağıda verilen satır komutu girilmeli ve çalıştırılmalıdır.

```
difTID(data[, 1:20], group = data[, 21], focal.name = 1)
```

Yukarıda verilen satır komutu incelendiğinde `difTID()` fonksiyonunun kullanıldığı ve bu fonksiyon içine üç argümanın girildiği görülmektedir. Bu argümanlardan ilki `data` argümanıdır, bu argümana veri seti (yalnızca maddeler) girilmelidir. İkinci argüman `group` argümanıdır ve bu argümana hangi iki grup arasında karşılaştırma yapılacağının bilgisinin girilmesi gerekmektedir. Son olarak üçüncü argüman `focal.name` argümanıdır. Bu argümana ise odak grubunun hangisi olduğunun bilgisi girilmelidir. Bu örnekte odak grup kadınlar olarak seçildiği için argümana 1 değeri girilmiştir. Satır komutu çalıştırıldığında maddelere ilişkin istatistikler ekrana gelecektir. Elde edilen madde istatistikleri Tablo 1'de verilmiştir.

Tablo 1.
Gerçekleştirilen DMF Analizi Sonucunda Elde Edilen Madde İstatistikleri

Madde	p_{j0}	p_{j1}	Δ_{j0}	Δ_{j1}	D_j
M1	.85	.45	8.85	13.50	-2.93*
M2	.85	.83	8.85	9.18	0.20
M3	.75	.85	10.30	8.85	1.44
M4	.75	.80	10.30	9.63	0.87
M5	.77	.48	10.04	13.20	-1.89*
M6	.69	.64	11.02	11.57	-0.04
M7	.72	.68	10.67	11.13	0.04
M8	.57	.59	12.29	12.09	0.46
M9	.59	.59	12.09	12.09	0.32
M10	.43	.56	13.71	12.40	1.21
M11	.47	.33	13.30	14.76	-0.78
M12	.41	.38	13.91	14.22	0.03
M13	.42	.44	13.81	13.60	0.41
M14	.34	.35	14.65	14.54	0.31
M15	.27	.33	15.45	14.76	0.70
M16	.29	.22	15.21	16.09	-0.43
M17	.25	.26	15.70	15.57	0.28
M18	.20	.17	16.37	16.82	-0.16
M19	.24	.19	15.83	16.51	-0.31
M20	.12	.13	17.70	17.51	0.26

Not. R çıktısından farklı olarak, değerler ondalık kısmında yalnızca iki basamak kalacak şekilde yuvarlanarak verilmiştir.

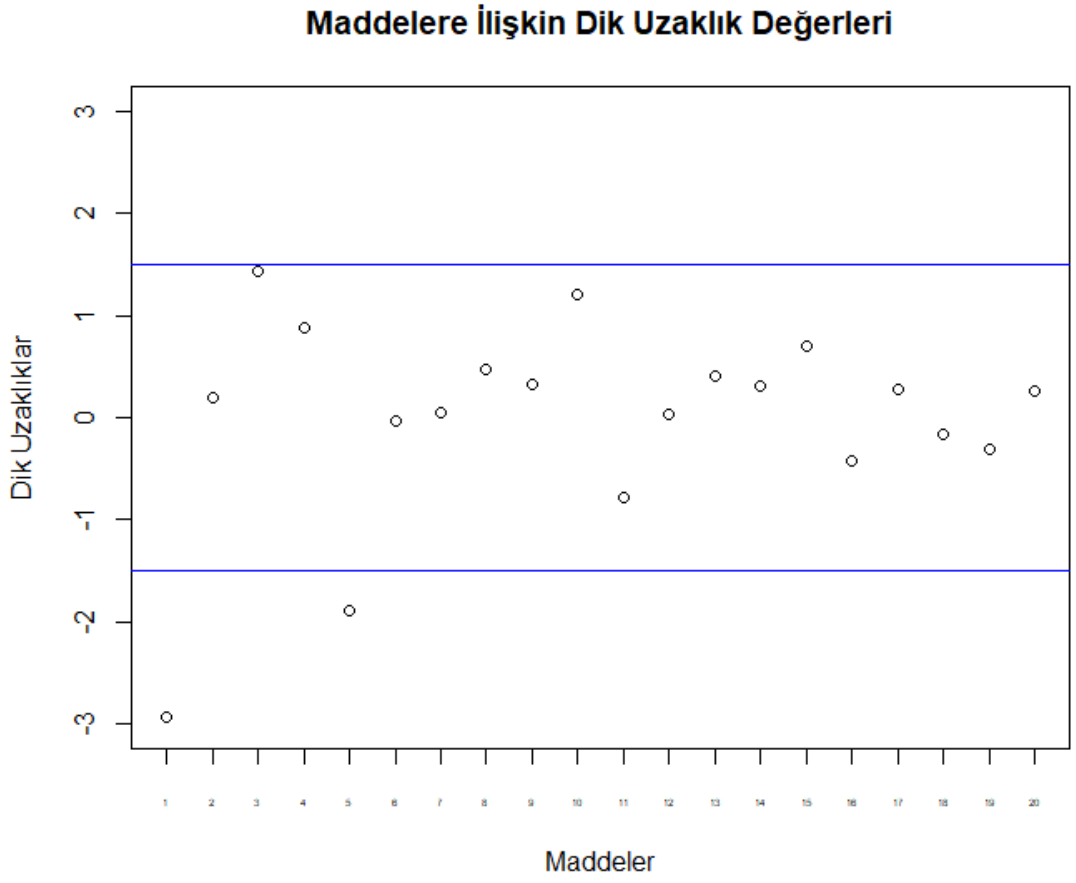
Tablo 1'de görüldüğü üzere Madde 1 (V1) ve Madde 5'te (V5) diğerlerinden farklı bir durum söz konusudur. Angoff'un DMG yöntemine göre bu maddeler DMF'li olarak işaretlenmiştir. R çıktısında (***) işaretlemesi, maddede gözlenen DMF'nin miktarının 1,5 eşik değerinin üzerinde olduğunu

göstermektedir. Çıktılar (analizi gerçekleştirdikten sonra ekrana gelecek) incelendiğinde ilk olarak çıktının en başında yöntemde sadeleştirme (purification) kullanılmadığının bilgisi verilmektedir. Hemen altında uç oranların (extreme proportions) 0,001 ile 0,999 arasına sıkıştırıldığının bilgisi verilmiştir. Bunun nedeni, ilgili maddeyi odak ya da referans gruptaki tüm bireyler doğru (bu durumda ilgili madde için p değeri 1 olacaktır) ya da yanlış (bu durumda ilgili madde için p değeri 0 olacaktır) yanıtladığında hesaplamaların gerçekleştirilemeyecek olmasından dolayı 0 değerinin 0,001'e, 1 değerinin de 0,999'a eşitleneceğinin bilgisinin verilmesidir. Bu çalışmadaki örnek uygulamada bu sıkıştırmaya ihtiyaç duyulacak bir madde bulunmamaktadır. Maddelere ilişkin istatistiklerin verildiği bölümde Prop.Ref (p_{j0}) sütununda basitçe referans grubunda ilgili maddeyi doğru yanıtlama oranı; Prop.Foc (p_{j1}) sütununda ise odak grubunda ilgili maddeyi doğru yanıtlama oranı yer almaktadır. Delta.Ref (Δ_{j0}) ve Delta.Foc (Δ_{j1}) sütunlarında sırasıyla referans ve odak gruplarında elde edilen doğru yanıt oranlarının (p_{jg}'lerin) delta değerlerine dönüştürülmüş değerleri yer almaktadır. Dist. (D_j) sütununda ise bu çalışmada da etraflıca bahsedilen dik uzaklık değerleri yer almaktadır. Zaten, bir maddenin DMF'li olup olmadığına bu sütundaki değerlere bakılarak karar verilmektedir. Olağan ayarlara göre Dist. sütununda [-1,5; 1,5] aralığının dışında bir değer bulunuyorsa bu satırdaki maddenin DMF gösterdiği söylenmektedir. Çıktılarda a (1.0546) ve b (0.9496) ile gösterilen parametreler, ana eksen doğrusunun denklemindeki katsayılardır. Delta noktaları ile bir grafiksel gösterimi tercih edenler için bu değerler kullanılarak bir doğru çizilebilir. Detection threshold kısmında 1.5 yazmaktadır. Bu `difTID()` fonksiyonunun olağan ayarlarında eşik değerinin 1,5 olarak seçildiğinin bir göstergesidir. İstenildiği takdirde fonksiyonun içine `thrTID` argümanı eklenerek bu değer değiştirilebilir. Son olarak, yanında üç yıldız işareti bulunan maddelerin listesi sunulmaktadır. Bu örnek uygulamadaki çıktıya göre V1 ve V5 kodlu maddeler DMF'li olarak işaretlenmiş maddelerdir.

3. Grafiksel gösterim. İsteyenler için '`difR`' paketi, DMF belirleme sürecinde elde edilen istatistikler ile bir grafik oluşturmaya da izin vermektedir. `difTID()` fonksiyonu ile elde edilen sonuçlara göre grafik çizdirmek istendiğinde R Düzenleyici penceresine aşağıdaki komut satırı girilmeli ve çalıştırılmalıdır.

```
plot(difTID(data[, 1:20], group = data[, 21], focal.name = 1))
```

Yukarıdaki satır komutu çalıştırıldığında, Grafik 1'deki gibi bir grafik elde edilecektir.

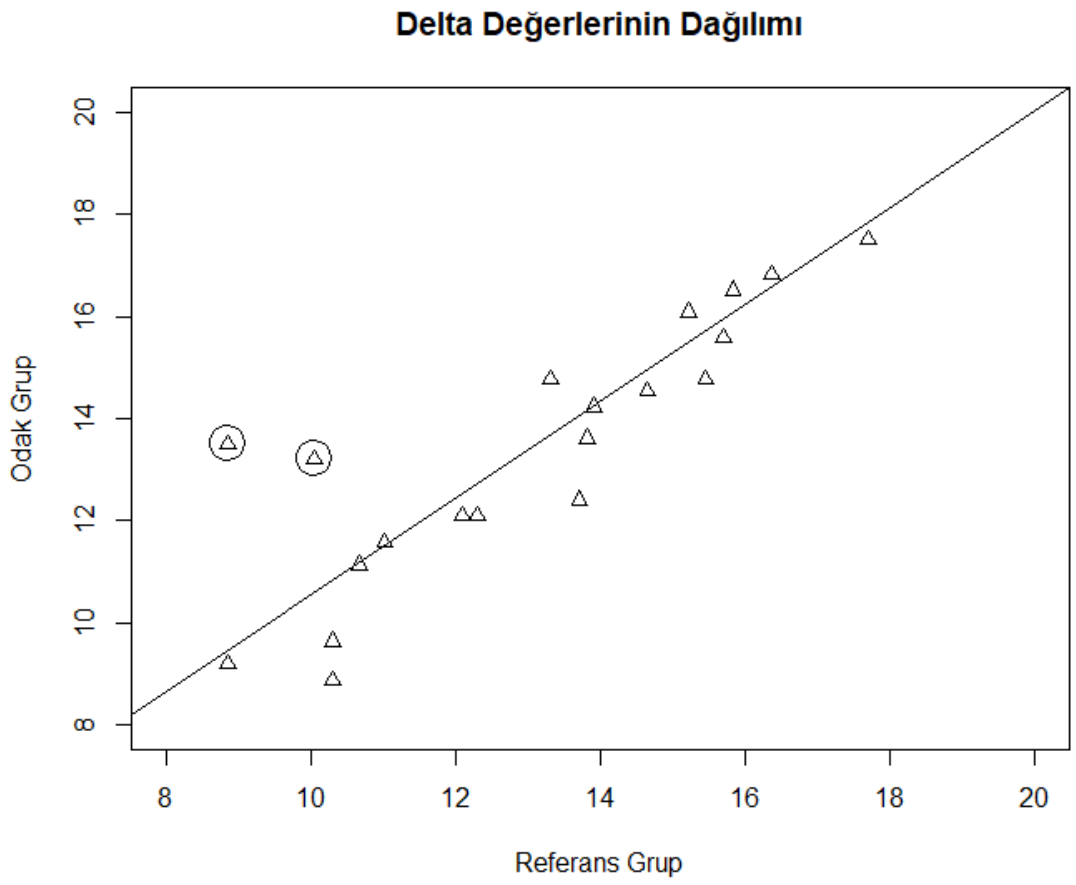


Grafik 1. *Dik Uzaklıklara İlişkin Grafik*

Grafik 1 incelendiğinde, yatay eksende madde numaralarının, dikey eksende ise çıktıdaki Dist. değerlerinin bulunduğu görülecektir. Grafik içinde bulunan iki yatay çizginin dikey eksende yer alan -1,5 ve +1,5 değerlerine karşılık geldiği fark edilecektir. Bunlar eşik değerleridir (threshold). Bu çizgilerin dışında kalan maddelere ait numaralar kırmızı renk ile işaretlenmektedir (R çıktısında). Ayrıca `plot()` fonksiyonunun içine `plot = "delta"` argümanı girilerek farklı bir grafik elde edilebilir.

```
plot(difTID(data[, 1:20], group = data[, 21], focal.name = 1), plot = "delta")
```

Yukarıdaki satır komutu çalıştırıldığında Grafik 2'deki gibi bir grafik elde edilecektir.



Grafik 2. *Odak ve Referans Gruptaki Delta Noktalarına İlişkin Saçılım Grafiği*

Grafik 2 incelendiğinde görülecektir ki, üçgen şekilleri ile gösterilen noktalar delta noktalarıdır ve referans gruba ait delta değerleri yatayda, odak gruba ait delta değerleri dikeyde yer almaktadır. İki üçgen şekli, birer çember içinde gösterilmektedir. Bunlar örnek uygulamadaki 1 ve 5. maddelerdir. Ortadan geçen ana eksenden belirgin şekilde ayrılan bu iki madde DMF'li olarak işaretlenmiştir.

4. Sonuçların raporlanması. Araştırmanın sonuçları raporlanırken şeffaf olunmalıdır ve diğer araştırmacıların da aynı veriler ile aynı araştırmayı yürüttüklerinde aynı sonuçlara ulaşmaları konusunda gereksinim duyulacak bilgiler verilmelidir. Bu noktada DMF analizi yürüten bir araştırmacı, başta kayıp veriler ve uç değerler ile nasıl başa çıktığını belirtmeli, eğer ek istatistikler (güvenirlik katsayısı, boyut yapısı analizi, test ve madde istatistikleri gibi) hesapladı ise paylaşmalı, ardından kullandığı DMF belirleme yöntemin sonuçlarını detaylıca paylaşmalıdır. Bu bağlamda, madde numaralarının, madde güçlüklerinin, delta değerlerinin, dik uzaklıkların ve kabul edilen eşik değerlere göre hangi maddelerin DMF'li olarak işaretlendiğinin belirtilmesi yerinde olacaktır.

4. TARTIŞMA ve SONUÇ

Madde/test yanlılığı konusu üzerine olan alan yazın incelendiğinde madde yanlılığının araştırılmasında yürütülecek istatistiksel süreçte (DMF analizi) Angoff'un DMG yöntemi, bilinen en eski yöntemlerden biri olarak öncü bir role sahiptir (Shepard vd., 1985). Bu yöntemden sonra pek çok DMF belirleme yöntemi geliştirilmiştir. Her bir yöntemin kendi içinde güçlü ve sınırlı yanları bulunmaktadır. Farklı yöntemlerle analizler gerçekleştirildiğinde farklı sonuçlar alınması olası olduğu için araştırmacıların yöntemlerin detaylarına hâkim olması önemlidir. Angoff'un yöntemi bu noktada, özellikle kolay anlaşılan algoritması, hesaplamaları ve görsel yorumlamaya olanak sağlaması ile öne çıkmaktadır (Angoff, 1982; Osterlind, 1983).

Yöntemin bir diğer önemli avantajı ise görece küçük örneklemlerden elde edilen verilerle analizlerin yapılabilmesidir (Muñiz vd., 2001). Yöntem, bu güçlü yönleriyle günümüzde kullanılmaktadır. Buna karşın, yöntemin pek çok sınırlı yönü de bulunmaktadır ve DMF analizi için DMF belirleme yöntemine karar vermeden önce araştırmacıların bu yönleri değerlendirmesi önerilmektedir.

Bu yöntem ile 1-0 şeklinde ikili puanlanmış maddelerden elde edilen veri setleri yine yalnızca iki grup üzerinden analiz edilebilmektedir. Çoklu puanlanan maddelerden elde edilen veri setlerinin analizinde ya da ikiden fazla grubun eş zamanlı karşılaştırılması gerektiği durumlarda bu yöntem kullanılamamaktadır. Bunun yerine genelleştirilmiş diğer DMF belirleme yöntemleri tercih edilmelidir. Yöntemin, günümüzdeki madde yanlılığı / DMF inceleme yaklaşımı açısından en önemli sınırlılığı ise yalnızca madde güçlüklerine odaklanıyor oluşudur. Bu durumda bir testi yanıtlayan gruplar arasında ölçülen özellik açısından farklılık söz konusu iken, yani gerçek bir fark bulunuyorken, Angoff'un yönteminin bu testteki ilgili maddeyi ya da maddeleri DMF'li olarak işaretleyecektir (Lord, 1977; Rudner, 1978). Oysa bu fark gerçek farktır (item impact). Bu önemli sınırlılıktan dolayı hem bu yönteme hem de bu yöntemin ardıllarına madde ayırt ediciliğini de hesaba katan algoritma geliştirmeleri yapılmıştır. Bunların yanı sıra bu yöntemde bir maddenin DMF içerdiğinin bir göstergesi olarak bir eşik değer belirtilmemiştir. Analiz sonucunda elde edilen Δ değerlerinin yorumlanması araştırmacıya bırakılmıştır. Bu çalışmada gerçekleştirilen örnek uygulamadan da görüldüğü üzere -2,93 ile 1,44 arasında değerler elde edilmiş ve bu değerlerden -1,5 ile 1,5 aralığı dışında kalanlar DMF'li olarak işaretlenmiştir. Bu eşik değer, bu çalışmada bu şekilde seçilmiş olup farklı çalışmalarda farklı şekillerde seçilebilmektedir. Bu nedenle, üzerinde fikir birliğine varılan bir eşik değere gereksinim duyulmaktadır.

Genel itibarıyla, avantaj ve dezavantajları değerlendirildiğinde, yöntemin bir testteki potansiyel olarak yanlı maddelerin belirlenmesinde bir öngörü sağlaması amacıyla sınırlı bir kullanımı (küçük örneklemlerle çalışılırken ya da görsel ve kolay yorumlamalara gereksinim duyulduğunda) düşünülebilir. Özellikle, karşılaştırılan grupların ölçülen özellik bakımından ortalamalarının eşit veya benzer olduğu durumda kullanımı daha uygun görünmektedir (Shepard vd., 1981). Bunun yanı sıra bir başka DMF belirleme yöntemi ile birlikte kullanımı ve ortak sonuçların raporlanması düşünülebilir.

Angoff'un DMG yönteminin DMF çalışmalarına bir ivme kazandırdığının, eksikliklerinin giderilerek farklı yöntemlerin geliştirilmesine öncü olduğunun altının çizilmesinde yarar vardır. Ayrıca, Angoff'un DMG yönteminin, madde/test yanlılığı gibi karmaşık ve derinlemesine incelemelere gereksinim duyulan konularda, özellikle DMF analizinin temelinin (mantığının) anlaşılmasında kullanılabilecek en iyi yöntemlerden biri olduğu söylenebilir. Özellikle ülkemizde ölçme ve değerlendirme, psikometri alanlarında lisansüstü derslerde madde ve test yanlılığı konularının daha somutlaştırılabilmesi adına daha detaylı olarak ele alınabilecek bir yöntem olan DMG yöntemi bu amaca hizmet etmesi için bu makalede tüm yönleriyle ele alınmıştır.

Bu çalışmada, yöntemin tüm yönleriyle ele alınmasının yanı sıra, sosyal bilimler alanındaki araştırmalarda da kullanılan R yazılımı aracılığı ile de uygulaması gerçekleştirilmiştir. Uygulamada `difR` paketi yardımıyla Angoff'un DMG yöntemi ile DMF analizinin adımları detaylı bir şekilde paylaşılmıştır. Makalenin gövdesinde verilen adımlar takip edilerek analiz kolay bir şekilde gerçekleştirilebilir. Gövdede satır satır açıklanan ve çalıştırılan satır komutları Ek-2'de sunulmuştur. Ayrıca, algoritması ve hesaplamaları oldukça basit olan Angoff'un DMG yöntemi ile DMF belirleme, R'daki temel fonksiyonlar ve argümanlar ile de gerçekleştirilebilir. Bunun için makale yazarları tarafından hazırlanan örnek satır komutları Ek-3'te verilmiştir. Bu çalışmada paylaşılan satır komutları ile diğer araştırmacıların da benzer adımları kullanarak tek başına bir DMF analizi gerçekleştirebilmesi mümkün kılınmaya çalışılmıştır.

Kaynakça / Reference

Aituariagbon, K. E., & Osarumwense, H. J. (2022). Non-parametric method of detecting differential item functioning in Senior School Certificate Examination (SSCE) 2019 Economics multiple choice items. *Kashere Journal of Education*, *3*(1), 146-158. https://dx.doi.org/10.4314/kje.v3i1.19

American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME]. (1999). *Standards for educational and psychological testing*. American Educational Research Association.

American Psychological Association [APA] (1988). *Code of fair testing practices in education*. Washington, DC: Author.

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (9th Ed.). Prentice-Hall, Inc.

Angoff, W. H. (1972). A technique for the investigation of cultural differences [Paper presentation]. American Psychological Association Annual Meeting, Honolulu.

Angoff, W. H. (1975). The investigation of test bias in the absence of an outside criterion [Paper presentation]. NIE Conference on Test Bias, Washington, D.C.

Angoff, W. H. (1982). Use of difficulty and discrimination indices for detecting item bias. In R. A. Beck (Ed.), *Handbook of methods for detecting item bias* (pp. 96-116). Johns Hopkins University Press.

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3-23). Lawrence Erlbaum Associates.

Angoff, W. H., & Cook, L. L. (1988). *Equating the scores of the prueba de aptitud académica and the scholastic aptitude test* (Report No. 88-3). ETS Research Report Series. https://doi.org/10.1002/j.2330-8516.1988.tb00259.x

Angoff, W. H., & Ford, S. F. (1973). Item-race interaction on a test of scholastic aptitude. *Journal of Educational Measurement*, *10*(2), 95-105. https://doi.org/10.1111/j.1745-3984.1973.tb00787.x

Angoff, W. H., & Modu, C. C. (1973). *Equating the scales of the Prueba de Aptitud Académica and the Scholastic Aptitude Test* (Report No. CEEB-RR-3). College Entrance Examination Board.

Bezruczko, N., Schulz, E. M., Reynolds, A. J., Perlman, C. L. & Rice, W. K. (1989). *The stability of four methods for estimating item bias* (Report No. ED-392-823). Department of Research and Evaluation, Chicago Public Schools.

Binet, A., & Simon, T. (1905). New methods for the diagnosis of the intellectual level of subnormals. In H. H. Goddard (Ed.), *Development of intelligence in children (the Binet-Simon Scale)*. Williams & Wilkins.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Sage Publications.

Choi, S. W., Gibbons, L. E., & Crane, P. K. (2011). Lordif: An R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and monte carlo simulations. *Journal of Statistical Software*, *39*(8), 1-30. https://doi.org/10.18637/jss.v039.i08

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice*, *17*(1), 31-44. https://doi.org/10.1111/j.1745-3992.1998.tb00619.x

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

de Ruiter, L. E., & Bers, M. U. (2022). The Coding Stages Assessment: Development and validation of an instrument for assessing young children's proficiency in the ScratchJr programming language. *Computer Science Education*, *32*(4), 1-30. https://doi.org/10.1080/08993408.2021.1956216

Devine, P. J., & Raju, N. S. (1982). Extent of overlap among four item bias methods. *Educational and Psychological Measurement*, *42*(4), 1049-1066. https://doi.org/001316448204200412

Dodeen, H., & Johanson, G. A. (2003). An analysis of sex-related differential item functioning in attitude assessment. *Assessment & Evaluation in Higher Education*, *28*(2), 129-134. https://doi.org/10.1080/02602930301667

Dorans, N. J. (1989). Two new approaches to assessing differential item functioning: Standardization and the Mantel-Haenszel Method. *Applied Measurement in Education*, *2*(3), 217-233. https://doi.org/10.1207/s15324818ame0203_3

Dorans, N. J., & Kulick, E. (1983). *Assessing unexpected differential item performance of female candidates on SAT and TSWE forms administered in December 1977: An application of the standardization approach* (Report No. RR-83-9). ETS Research Report Series. https://doi.org/10.1002/j.2330-8516.1983.tb00009.x

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement*, *23*(4), 355-368. https://doi.org/10.1111/j.1745-3984.1986.tb00255.x

Dzul-Garcia, C., & Atar, B. (2020). Investigation of possible item bias on PISA 2015 science items across Chile, Costa Rica and Mexico. *Culture and Education*, *32*(3), 470-505. https://doi.org/10.1080/11356405.2020.1785158

Elosua, P., & Wells, C. S. (2013). Detecting DIF in polytomous items using MACS, IRT and ordinal logistic regression. *Psicológica*, *34*(2), 327-342.

Facon, B., & Nuchadee, M-L. (2010). An item analysis of Raven's Colored Progressive Matrices among participants with Down syndrome. *Research in Developmental Disabilities*, *31*(1), 243-249. https://doi.org/10.1016/j.ridd.2009.09.011

Farcomeni, A., Pittau, M. G., Viviani, S., & Zelli, R. (2022). A European measurement scale for material deprivation. *Research Square*, 1-32. https://doi.org/10.21203/rs.3.rs-2250804/v1

Finch, W. H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement*, *29*(4), 278-295. https://doi.org/10.1177/0146621605275728

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2012). *How to design and evaluate research in education* (8th Ed.). Mc-Graw Hill.

Gamerman, D., Gonçalves, F. B., & Soares, T. M. (2018). Differential item functioning. In W. J. van der Linden (Ed.), *Handbook of item response theory* (pp. 67-84). CRC Press.

Gao, Y., & Zhu, W. (2009). Identifying culturally sensitive physical activities using DIF analysis. *Medicine & Science in Sports & Exercise*, *41*(5), 416-417. http://dx.doi.org/10.1249/01.MSS.0000355818.07045.09

Gelin, M. N., Carleton, B. C., Smith, A. A., & Zumbo, B. D. (2004). The dimensionality and gender differential item functioning of the mini asthma quality of life questionnaire (MINIAQLQ). *Social Indicators Research*, *68*(1), 91-105. https://doi.org/10.1023/B:SOCI.0000025580.54702.90

Gómez-Benito, J., Sireci, S., Padilla, J.-L., Hidalgo, M. D., & Benítez, I. (2018). Differential item functioning: Beyond validity evidence based on internal structure. *Psicothema*, *30*(1), 104-109. http://doi.org/10.7334/psicothema2017.183

Gould, S. J. (1981). *The mismeasure of man*. W. W. Norton & Company.

Hauger, J. B., & Sireci, S. G. (2008). Detecting differential item functioning across examinees tested in their dominant language and examinees tested in a second language. *International Journal of Testing*, *8*(3), 237-250. http://dx.doi.org/10.1080/15305050802262183

Holland, P. W., & Thayer, D. T. (1988). Differential item functioning and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129-145). Lawrence Erlbaum Associates.

Holland, P. W., & Wainer, H. (Eds.). (1993). Differential item functioning: Theory and practice. Erlbaum Publishers.

Hunter, J. E. (1975). A critical analysis of the use of item means and item-test correlations to determine the presence or absence of content bias in achievement test items [Paper presentation]. National Institute of Education Conference on Test Bias, Annapolis, MD.

Ironson, G. H., & Subkoviak, M. J. (1979). A comparison of several methods of assessing item bias. *Journal of Educational Measurement*, *16*(4), 209-225. https://doi.org/10.1111/j.1745-3984.1979.tb00103.x

Iwata, N., Turner, R. J., & Lloyd, D. A. (2002). Race/ethnicity and depressive symptoms in community-dwelling young adults: A differential item functioning analysis. *Psychiatry Research*, *110*(3), 281-289. https://doi.org/10.1016/S0165-1781(02)00102-6

Jensen, A. R. (1973). *Educability and group differences*. Basic Books.

Jensen, A. R. (1976). Test bias and construct validity. *The Phi Delta Kappan*, *58*(4), 340-346.

Kelderman, H. (1989). Item bias detection using loglinear IRT. *Psychometrika*, *54*(4), 681-697. https://doi.org/10.1007/BF02296403

Korkmaz, M. (2006). Test ve ölçek geliştirmede yeni yaklaşımlar: Madde cevap kuramı kapsamında madde işlevsel farklılık (madde yanlılık) yöntemleri. *Türk Psikoloji Yazıları*, *9*(18), 63-80.

Li, Z., & Zumbo, B. D. (2009). Impact of differential item functioning on subsequent statistical conclusions based on observed test score data. *Psicológica*, *30*(2), 343-370.

Lord, F. M. (1977). A study of item bias using item characteristic curve theory. In N. H. Poortinga (Ed.), *Basic problems in cross-cultural psychology* (pp. 19-29). Swets and Zeitlinger.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum Associates, Inc.

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 Research Methods*, *42*, 847-862. https://doi.org/10.3758/BRM.42.3.847

Magis, D., & Facon, B. (2012). Angoff's delta method revisited: Improving DIF detection under small samples. *British Journal of Mathematical and Statistical Psychology*, *65*(2), 302-321. https://doi.org/10.1111/j.2044-8317.2011.02025.x

Magis, D., & Facon, B. (2014). deltaPlotR: An R package for differential item functioning analysis with Angoff's Delta Plot. *Journal of Statistical Software*, *59*(1), 1-19. https://doi.org/10.18637/jss.v059.c01

Mellenberg, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research*, *13*, 127-143. https://doi.org/10.1016/0883-0355(89)90002-5

Muñiz, J., Hambleton, R. K., & Xing, D. (2001). Small sample studies to detect flaws in item translations. *International Journal of Testing*, *1*(2), 115-135. https://doi.org/10.1207/S15327574IJT0102_2

Oosterhof, A. C., Atash, M. N., & Lassiter, K. L. (1984). Facilitating identification of item bias through use of delta plots. *Educational and Psychological Measurement*, *44*(3), 619-627. https://doi.org/10.1177/0013164484443009

Osterlind, S. J. (1983). *Test item bias*. Sage Publications.

Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning* (2nd Ed.). Sage Publications.

Ozarkan, H. B., Kucam, E. ve Demir, E. (2017). Merkezi Ortak Sınav Matematik alt testinde değişen madde fonksiyonunun görme engeli durumuna göre incelenmesi. *Curr Res Educ*, *3*(1), 24-34.

Penfield, R. D., & Camilli, G. (2007). Differential item functioning and item bias. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 26, pp. 125-167). Elsevier.

Pine, S. M. (1977). Applications of item response theory to the problem of test bias. In D. J. Weiss (Ed.), *Applications of computerized adaptive testing* (pp. 37-43). University of Minnesota, Psychometric Methods Program.

R Development Core Team. (2023). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, *53*(4), 495-502. https://doi.org/10.1007/BF02294403

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement*, *14*(2), 197- 207. https://doi.org/10.1177/014662169001400208

Raju, N. S., Drasgow, F., & Slinde, J. A. (1993). An empirical comparison of the area methods, Lord'S Chi-Square Test, and the Mantel-Haenszel technique for assessing differential item functioning. *Educational and Psychological Measurement*, *53*(2), 301-314. https://doi.org/10.1177/0013164493053002001

Revelle, W. (2023). psych: Procedures for Psychological, Psychometric, and Personality Research. Northwestern University, Evanston, Illinois. R package version 2.3.6, https://CRAN.R-project.org/package=psych.

Robin, F., Sireci, S. G., & Hambleton, R. K. (2003). Evaluating the equivalence of different language versions of a credentialing exam. *International Journal of Testing*, *3*(1), 1-20, https://doi.org/10.1207/S15327574IJT0301_1

Rudner, L. M. (1978). Using standard tests with the hearing impaired: The problems of item bias. *Volta Review*, *80*, 31-40.

Scarr, S., & Weinberg, R. A. (1976). IQ test performance of Black children adopted by White families. *American Psychologist*, *31*(10), 726-739. https://doi.org/10.1037/0003-066X.31.10.726

Scheuneman, J. (1979). A method of assessing bias in test items. *Journal of Educational Measurement*, *16*(3), 143-152. https://doi.org/10.1111/j.1745-3984.1979.tb00095.x

Seong, T-J., & Subkoviak, M. J. (1987). A comparative study of recently proposed item bias detection methods [Paper presentation]. Annual Meeting of the National Council on Measurement in Education, Washington, D.C.

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. https://doi.org/10.1007/BF02294572

Shepard, L. A., Camilli, G., & Averill, M. (1981). Comparison of procedures for detecting test-item bias with both internal and external ability criteria. *Journal of Educational Statistics*, *6*(4), 317-375. https://doi.org/10.3102/10769986006004317

Shepard, L. A., Camilli, G., & Williams, D. A. (1985). Validity of approximation techniques for detecting item bias. *Journal of Educational Measurement*, *22*(2), 77-105. https://doi.org/10.1111/j.1745-3984.1985.tb01050.x

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

Tat, O. ve Doğan, N. (2018). Uluslararası Bilgisayar ve Bilgi Teknolojileri Okuryazarlığı Testinin madde-birey dağılımı ve değişen madde fonksiyonu yönünden incelenmesi. *Gazi Üniversitesi Gazi Eğitim Fakültesi Dergisi*, *38*(3), 1207-1231. https://doi.org/10.17152/gefad.321630

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 147-172). Lawrence Erlbaum Associates, Inc.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67-113). Lawrence Erlbaum Associates, Inc.

Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology*, *16*(7), 433-451. https://doi.org/10.1037/h0073357

van der Flier, H., Mellenberg, G. J., Adér, H. J., & Wijn, M. (1984). An iterative item bias detection method. *Journal of Educational Measurement*, *21*(2), 131-145. https://doi.org/10.1111/j.1745-3984.1984.tb00225.x

Van Vo, D., & Csapó, B. (2023). Effects of multimedia on psychometric characteristics of cognitive tests: A comparison between technology-based and paper-based modalities. *Studies in Educational Evaluation*, *77*, 1-12. https://doi.org/10.1016/j.stueduc.2023.101254

Wainer, H. (1993). Model-based standardized measurement of an item's differential impact. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 123-135). Lawrence Erlbaum Associates.

Wainer, H., Bradlow, E., & Wang, X. (2010). Detecting DIF: Many paths to salvation. *Journal of Educational and Behavioral Statistics*, *35*(4), 489-493. https://doi.org/10.3102/1076998610376624

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

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, *4*(2), 223-233. http://dx.doi.org/10.1080/15434300701375832

Zwick, R., & Ercikan, K. (1989). Analysis of differential item functioning in the NAEP history assessment. *Journal of Educational Measurement*, *26*(1), 55-66. https://doi.org/10.1111/j.1745-3984.1989.tb00318.x

EKLER

Ek-1. Analiz edilecek simülasyon verinin hazır hale getirilmesi için kullanılan satır komutları

```
# 'psych' paketinin kurulması ve çalıştırılması (installing and running "psych" package)
install.packages("psych")
library(psych)

# veri üretimi için bir başlangıç değerinin girilmesi (setting seed for data generation)
set.seed(12)

# Rasch modele uygun bir veri setinin oluşturulması (generating data)
data_rasch <- sim.rasch(nvar = 20, n = 200, low = -2, high = 2,
                                d = NULL, a = 1, mu = 0, sd = 1)

# kullanılacak olan veri setinin görüntülenmesi (viewing data)
data_rasch$items

# cinsiyet gruplarının oluşturulması (creating gender groups)
x <- c(1, 0)
gender <- c(rep(1, times = 100), rep(0, times = 100))

# cinsiyet gruplarının veri setine eklenmesi (1 = kadın, 0 = erkek) (adding gender groups to
the data)
data <- cbind(data_rasch$items, gender)

# kadınlarda Madde 1 ve Madde 5'e birkaç 0 değerinin atanması (manipulating data – Item1 &
Item5)
data[1:50, 1] <- rep(0, times = 50)
data[1:30, 5] <- rep(0, times = 30)

# analiz edilecek son veri setinin görüntülenmesi (viewing final data)
data
```

Ek-2. "difR" paketi yardımıyla DMF analizi

```
# 'difR' paketinin kurulması ve çalıştırılması (installing and running "difR" package)
install.packages("difR")
library(difR)

# veri setinin analiz öncesinde incelenmesi (viewing data)
## not: bu çalışmada veri seti nesnesinin adı 'data'dır. Sizde bu nesnenin adı ne ise
fonksiyonların içine o girilmelidir.
dim(data)
head(data)

# Angoff'un DMG yöntemi ile analizin gerçekleştirilmesi (analyzing the data by Angoff's Method)
## not: 1:20 ve 21 sayıları değişkenlerin sütun numaralarına karşılık gelmektedir. Buna göre
1'den 20'ye kadar olan maddeler DMF analizine dahil edilecektir. Cinsiyet değişkeni (1-0) veri
setinin 21. sütununda yer almaktadır ve kadınlar (1) odak gruptur.
difTID(data[, 1:20], group = data[, 21], focal.name = 1)

# grafiksel gösterim (dik uzaklıklar) (perpendicular distances – graph)
plot(difTID(data[, 1:20], group = data[, 21], focal.name = 1))

# grafiksel gösterim (delta noktaları) (delta points – graph)
plot(difTID(data[, 1:20], group = data[, 21], focal.name = 1), plot = "delta")
```

Ek-3. R'daki temel fonksiyon ve argümanlarla DMF analizi

```
# veri setinin incelenmesi (viewing data)
data

# odak ve referans grubun ayrılması (separating the focus and the reference groups)
women <- subset(data, gender == 1)
men <- subset(data, gender == 0)

# odak ve referans grupta maddelere ilişkin p değerlerinin hesaplanması (calculating p values)
FocColMeans <- colMeans(women[, 1:20])
RefColMeans <- colMeans(men[, 1:20])

# odak ve referans gruptaki delta değerlerinin hesaplanması (calculating delta values)
DeltaFoc <- 4*qnorm(1 - FocColMeans) + 13
DeltaRef <- 4*qnorm(1 - RefColMeans) + 13

# gereksinim duyulacak istatistiklerin hesaplanması (calculating the needed statistics)
x1 <- mean(DeltaFoc)
x0 <- mean(DeltaRef)
VarFoc <- var(DeltaFoc)
VarRef <- var(DeltaRef)
Covariance <- cov(DeltaFoc, DeltaRef)

# b, a (ana eksen doğrusuna ait katsayılar) ve uzaklıkların hesaplanması (calculating
perpendicular distances, a and b)
b <- (VarFoc - VarRef + sqrt((VarFoc - VarRef)^2 + 4*Covariance^2)) / (2*Covariance)
a <- x1 - b*x0
D <- (b*DeltaRef + a - DeltaFoc) / sqrt(b^2 + 1)

# eşik değere göre DMF'li olarak işaretlenecek maddeleri ayırt edecek bir işaretlemenin
girilmesi (flagging DIF items)
Treshold <- D < -1.5 | D > 1.5
Flag <- ifelse(Treshold == "TRUE", "***", "-")

# istatistikler (stats)
Stats <- cbind(Women_p = FocColMeans, Men_p = RefColMeans, Women_Delta = round(DeltaFoc, 4),
Men_Delta = round(DeltaRef, 4), Distance = round(D, 4))
cbind(as.data.frame(Stats), Flag)

# grafiksel gösterim (graph)
plot(DeltaRef, DeltaFoc, pch = 4)
abline(a, b, col = "blue")
```

EXTENDED ABSTRACT

1. INTRODUCTION

The situation where the scores obtained from a measurement instrument depend on certain characteristics other than the measured trait of the individuals responding to the measurement instrument reveals the risks of test and item bias. The variations in response behaviors of individuals from different groups in answering the items of a test, especially emphasize the need for item bias studies (Mellenberg, 1989; van der Flier et al., 1984). If respondents from different groups (race, gender, etc.) show variations in their responses to test items, the effects of this situation must be thoroughly examined. Item bias, briefly, refers to the favoring or disfavoring of a specific subgroup by an item (Osterlind & Everson, 2009). Such bias poses a threat to the validity of the scores obtained from the test and causes systematic error in measurement (Clauser & Mazor, 1998; Osterlind, 1983).

Practical applications for bias studies on tests and items began to accelerate and become widespread in the late 1960s. The number of studies in this scope has been increasing, especially with the recognition of their effects in psychology, social sciences, and educational sciences (Gómez-Benito et al., 2018). Many statistical techniques commonly used today to identify items with potential bias in a test started to be developed in the 1970s. Penfield and Camilli (2007) explain these statistical techniques, taking their historical backgrounds into account. Thanks to valuable contributions such as the comprehensive work of Holland and Thayer (1988), significant progress has been made in identifying items with potential bias and detecting differential item functioning (DIF) through bias detection studies.

Despite its historical significance, Angoff's Transformed Item Difficulties (TID) method has often been criticized due to its limitations. The purpose of this study is to closely examine the algorithm of Angoff's TID method, which is frequently criticized in DIF studies but still used in some studies today and question its usability in research. To achieve this goal, the following questions were addressed:

1) What are the strengths and limitations of Angoff's TID method in DIF detection?

2) Based on the algorithm of the method, in which studies/fields can the method be applied?

2. METHOD

2.1. Algorithm of Angoff's TID Method

In accordance with Angoff's method, calculations for detecting an item exhibiting Differential Item Functioning (DIF) can be performed manually or using a basic spreadsheet software. In this section, the step-by-step algorithm of Angoff's Transformed Item Difficulties (TID) method will be explained.

1. Firstly, the groups to be compared must be identified. One of these groups will be the focal group, and the other will be the reference group. For example, consider individuals who responded to the items of a test along with their gender information collected after a test administration. If the focal group is selected as females and the reference group as males, a binary assignment of 1 to female individuals and 0 to male individuals will be made.

2. For each item, the item difficulty index should be calculated separately for each group (focal and reference). The item difficulty indices, denoted by p_{jg}, should be computed using the subscript g to represent the group, where g is an index for groups.

3. The item difficulty indices should be converted to standardized z-scores. In this step, the obtained p_{jg} values are transformed into values corresponding to the z-scores table. For instance, the values 0.95 and 0.975 will correspond to -1.64 and -1.96, respectively. The scores obtained through the transformation of p_{jg} values for different groups are denoted as z_{jg}.

4. The obtained z_{jg} scores should be transformed into Δ values. In this step, a linear transformation recommended by Angoff is applied. Accordingly, each z-score in each group is transformed into a Δ value using the equation $\Delta_{jg} = 4z_{jg} + 13$ (where the mean is 13, and the standard deviation is 4).

5. The obtained Δ values $(\Delta_{j0}, \Delta_{j1})$ for two separate groups are examined through a scatter plot. The delta values for two distinct groups can be visually inspected in a distribution graph where the reference group is on the horizontal axis and the focal group is on the vertical axis. This graph is specifically referred to as a delta plot.

6. In addition to graphical examination, calculating the perpendicular distances to the major axis of the delta points is necessary to detect whether items exhibit DIF. The major axis of the ellipse is expressed by the equation $\Delta_{j1} = a + b\Delta_{j0}$. At this point, the intersection (a) and slope (b) parameters need to be calculated.

The calculations for the intersection and slope parameters are conducted using the equations:

$$b = \frac{s_1^2 - s_0^2 + \sqrt{(s_1^2 - s_0^2)^2 + 4s_{01}^2}}{2s_{01}}$$

$$a = \bar{x}_1 - b\bar{x}_0$$

Here, $\bar{x}_1$ and $\bar{x}_0$ represent the means of the focal and reference groups, s_1^2 and s_0^2 represent the variances of the focal and reference groups, and s_{01} represents the covariance.

The perpendicular distances (D_j) expressing the distances of delta points to the major axis are calculated as follows:

$$D_j = \frac{b\Delta_{j0} + a - \Delta_{j1}}{\sqrt{b^2 + 1}}$$

As understood from the equation, the distances of points to a line are being calculated. Large distance values indicate that the delta points deviate from the major axis, suggesting the potential for DIF in the item.

Various recommendations exist regarding the number of perpendicular distances at which suspicion of DIF arises for an item. These recommendations serve a technical classification purpose (whether the item exhibits DIF or not). The suggested threshold values do not refer to a statistical significance level; instead, they can only be considered as an effect size. In the literature, a commonly used threshold value indicating DIF at the C level, which is frequently employed in interpreting the Mantel-Haenszel method, is 1.5 (Zwick & Ercikan, 1989). This threshold value is also used in Angoff's TID method.

3. FINDINGS, DISCUSSION AND RESULTS

3.1. Strengths of Angoff's TID Method

The simplicity of the algorithm is one of the strengths of Angoff's TID method. The sample size in DIF detection studies has the power to influence item statistics. In some methods, working with relatively large samples is often considered a necessity. However, Angoff's TID method allows for the conduct of studies with small samples (Muñiz et al., 2001). The mathematical operations involved in the method are quite straightforward. Calculating delta values, major axis, and perpendicular distances in the algorithm is easy, and delta values (points) can be visualized with a graph. Interpretations can be easily made after graphical examination. For these reasons, despite its limitations, Angoff's TID method can still be used in research.

3.2. Limitations of Angoff's TID Method

Angoff's TID method focuses on the correct response rates for an item across groups. This method does not consider item discrimination. If an item's discrimination index vary across groups (non-uniform DIF), this method may not be suitable for detecting DIF. This is one of the significant limitations of the method.

In this method, a mathematical issue can arise when every individual in a group answers an item correctly or incorrectly. This is because in such cases, the p_j value will be either 1 or 0, leading to infinite delta values. To overcome this limitation, Angoff and Ford (1973) proposed constraining item difficulties to the range [0.05; 0.95].

One significant limitation of the method is the presence of item impact and item discrimination. As expressed by Hunter (1975), Lord (1977), and Rudner (1978), there may be differences among responders in terms of the measured attribute. In such cases, average item difficulties may vary as groups have different mean proficiency levels. Even though item impact is present, the method would mark the item as exhibiting DIF. Angoff acknowledged this issue later on (Angoff, 1993). In response to criticisms that the method focuses solely on item difficulties, Angoff (1982) made an adjustment to include item discriminations in the calculations, resulting in the development of the Revised Transformed Item Difficulties Method (Revised TID Method). However, this method was also deemed insufficient for item bias investigations (Seong & Subkoviak, 1987).

Another limitation of Angoff's TID method is related to threshold values. In this method, the decision of whether an item exhibits DIF is based on the perpendicular distance of the item's delta value to the correct response. The detection of how much distance qualifies as indicating DIF is a subject of debate. Although a common threshold value is often set at 1.5 in the literature, Angoff did not explicitly recommend a specific threshold.

Furthermore, Angoff's TID method can only be applied to binary scored items (dichotomous). It cannot analyze items scored in more than two categories (polytomous). Additionally, this method allows the comparison of only two groups. When comparing more than two groups, it is necessary to use generalized methods for detecting DIF.

3.3. In Which Research Contexts Can It Be Used?

Angoff's TID method is designed to assess whether items in a test provide an advantage in terms of item difficulty for individuals in different groups. In essence, if items in a test are answered more easily by individuals in one group compared to individuals in another group, the method provides an indication that the items may exhibit DIF. However, this approach is somewhat limited from the perspective of contemporary DIF analysis. This is because items have parameters beyond difficulty parameters, and without considering these, incorrect interpretations may occur. Due to the theoretical limitation of focusing solely on item difficulties, this method can yield more functional results when the total score averages for the measured attribute are equal or very close between the two compared groups (Shepard et al., 1981). Additionally, according to this method, if certain items are more discriminating than others, the method may label them as exhibiting DIF.

In contrast to later-developed DIF detection methods following Angoff's TID method, it is observed that ability matching is conducted when comparing groups. This is because comparing based on group averages may lead to erroneous comparisons of individuals in lower and upper achievement groups. Angoff's TID method has a limitation due to the absence of ability matching in DIF analysis and the comparisons being made based on group averages.

In summary, the use of Angoff's TID method for DIF analysis should not stand alone in significant research. However, limited use can be considered when working with small samples and in situations where significant differentiation in the measured attribute is not expected between groups, aiming to provide an indication.

4. DISCUSSION AND RESULTS

When examining the literature on item/test bias, Angoff's TID method stands out as one of the pioneering methods in the statistical process for investigating item bias (Shepard et al., 1985). This method holds a prominent role, especially due to its easily understandable algorithm and calculations. Evaluating its advantages and disadvantages, the method can be considered for limited use with the aim of providing an indication of potentially biased items in a test. Its usage appears more suitable, particularly when the averages of the compared groups are equal or similar concerning the measured attribute.

It is worth emphasizing that Angoff's TID method has played a pivotal role in accelerating DIF studies and has paved the way for the development of different methods by addressing its shortcomings. Additionally, in complex and in-depth investigations such as item/test bias, especially in understanding the logic of DIF analysis, Angoff's TID method can be considered one of the best methods available. Particularly in our country, where the topics of item and test bias in postgraduate courses in measurement and evaluation, and psychometrics could be more concretely addressed, TID method, being a detailed method, has been comprehensively discussed in this article to serve this purpose.

In this study, not only has the method been thoroughly discussed, but its application has also been carried out using the R software. In the application, the steps of DIF analysis with Angoff's TID method using the difR package have been detailed. The analysis can be easily conducted using the line commands provided in APP-2. Additionally, detecting DIF with Angoff's TID method, which has a simple algorithm and calculations, can also be performed in R using basic functions and arguments. Sample line commands prepared by the authors for this purpose are provided in APP-3. Through the shared line commands in this study, an attempt has been made to enable other researchers to independently conduct a DIF analysis by following similar steps.

ARAŞTIRMANIN ETİK İZNİ

Bu çalışmada "Yükseköğretim Kurumları Bilimsel Araştırma ve Yayın Etiği Yönergesi" kapsamında uyulması gerektiği 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: Ankara Üniversitesi Rektörlüğü Etik Kurulu Başkanlığı

Etik değerlendirme kararının tarihi: 03.01.2024

Etik değerlendirme belgesi sayı numarası: 85434274-050.04.04 / 1219852

ARAŞTIRMACILARIN KATKI ORANI

MG: Araştırmanın tasarlanması, alan yazını taraması, veri analizi, raporlaştırma.

ED: Yöntemin belirlenmesi, danışmanlık.

ÇATIŞMA BEYANI

Araştırmada herhangi bir kişi ya da kurum ile finansal veya kişisel yönden bağlantı ya da çıkar çatışması bulunmamaktadır.