Yayın Geliş Tarihi: 29.09.2022
Yayına Kabul Tarihi: 23.02.2023
Online Yayın Tarihi: 15.03.2023
http://dx.doi.org/10.16953/deusosbil.1181867

Dokuz Eylül Üniversitesi
Sosyal Bilimler Enstitüsü Dergisi
Cilt: 25, Sayı: 1, Yıl: 2023 Sayfa: 227-245
E-ISSN: 1308-0911

Araştırma Makalesi

PUANLAYICILAR ARASI UYUMUN FARKLI ÖLÇEKLEME TÜRLERİ, PUANLAYICI SAYISI VE PUANLANAN SAYISI AÇISINDAN İNCELENMESİ

Yılmaz Orhun GÜRLÜK[*]

Mediha KORKMAZ[**]

Gizem CÖMERT[***]

Ö. Emre C. ALAGÖZ[****]

Öz

Bu araştırmada klasik kuramlara göre puanlayıcılar arası uyum katsayılarını karşılaştırmak amaçlanmıştır. Farklı ölçekleme türlerine göre elde edilen katsayılar üzerinden hesaplanan değerler arasındaki farka odaklanılmış ve ölçekleme türüne karar vermenin önemi ortaya konmuştur. Puanlanan ve puanlayıcı sayısının değişmesinin değerleri etkileyip etkilemediğine bakılmış ve genellenebilirlik kuramının optimizasyon

Bu makale için önerilen kaynak gösterimi (APA 6. Sürüm):

Gürlük, Y. O., Korkmaz, M., Cömert, G. & Alagöz, Ö. E. C. (2023). Puanlayıcılar arası uyumun farklı ölçekleme türleri, puanlayıcı sayısı ve puanlanan sayısı açısından incelenmesi. *Dokuz Eylül Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 25* (1), 227-245.

[*] Doktora Öğrencisi, Doktorant, Ege Üniversitesi, Sosyal Bilimler Enstitüsü, Psikoloji / Psikometri Anabilim Dalı, ORCID: 0000-0002-1134-3776, yilmazorhungurluk@gmail.com.

[**] Doç. Dr., Ege Üniversitesi, Edebiyat Fakültesi, Psikoloji Bölümü, Psikometri Anabilim Dalı, ORCID: 0000-0001-6504-5822, medihakrkmz@gmail.com.

[***] Doktora Öğrencisi, Doktorant, Ege Üniversitesi, Sosyal Bilimler Enstitüsü, Psikoloji / Psikometri Anabilim Dalı, ORCID: 0000-0001-7555-6378, cmrtgizem@gmail.com.

[****] Doktora Öğrencisi, Doktorant, Mannheim Üniversitesi, Psikoloji Bölümü, Ölçme ve Değerlendirme Anabilim Dalı, ORCID: 0000-0003-3305-6564, oemrecalagoz@gmail.com.

Araştırmacı Ömer Emre Can Alagöz, Deutsche Forschungsgemeinschaft (DFG, Alman Araştırma Kurumu) - GRK 2277 "Statistical Modeling in Psychology" (SMiP; Psikolojide İstatistiksel Modelleme) nolu proje tarafından desteklenmektedir.

Araştırmada analiz aşamasında kullanılan veriler, Korkmaz vd.,(2015-2019) tarafından yürütülen Bender Gestalt-II Testinin Koppitz II ve Bender Gestalt II Puanlama Sistemlerine Göre (4-7 ve 8-18 Yaş) Ön Norm Çalışması (Ege Üniversitesi Bilimsel Araştırma Projesi, Proje No: 15-EDB-021) kapsamında, İzmir İl Milli Eğitim Müdürlüğünden Etik Kurul onayına sahiptir (30/12/2015 tarihli 12018877-604.01.02-E.13519131).

analizi kullanılarak puanlayıcılar arası uyum için kullanılacak en uygun örneklem büyüklüğü hesaplanmıştır. Araştırmada toplamda 35 çocuğa Bender Görsel Motor Gestalt II testinin yaş gruplarında ortak olan 9 kopyalama kartı uygulanmış ve alınan ölçümler toplamda 8 puanlayıcı tarafından birbirlerine kör olarak değerlendirilmiştir. Sonuçlara göre en yüksek uyum değeri sınıf içi korelasyon katsayısında hesaplanmış ve bu değeri sırasıyla Krippendorff alfa, Fleiss kappa ve Cohen kappa takip etmiştir. Hem puanlanan hem de puanlayıcı sayısı azaldıkça uyum değerlerinin düşme eğiliminde olduğu tespit edilmiştir. Öte yandan kartların zorluk düzeyinin anlamlı bir etkisi olmadığı saptanmıştır. Genellenebilirlik katsayılarının yüksek çıkması testin puanlayıcılar tarafından güvenilir şekilde puanlandığını göstermiştir. Optimizasyon analizi incelendiğinde bu test için en uygun örneklem büyüklüğünün 50 olduğu görülmüştür. Katılımcı sayısının 50'den fazla olması ise uyumu arttırmamıştır.

Anahtar Kelimeler: *Puanlayıcılar Arası Uyum, Ölçekleme Türü, Sınıf İçi Korelasyon, Kappa, Krippendorff Alfa, Genellenebilirlik Kuramı.*

EXAMINING THE INTERRATER RELIABILITY ACCORDING TO DIFFERENT SCALING TYPES, NUMBER OF RATERS AND NUMBER OF RATED SUBJECTS

Abstract

In this study, it was aimed to compare the coefficients of interrater agreement according to classical statistic theories. The difference between the calculated agreement coefficients according to different scaling types has been focused and the importance of deciding on the scaling type has been revealed. It was examined whether the change in the number of raters and rateds affected the values, and the most appropriate sample size to be used for the interrater agreement was calculated by using the optimization analysis of the generalizability theory. In the study, 9 cards of the Bender-Gestalt motor skill test, which can be seen by everyone, were applied to 35 children in total, and the measurements were evaluated by 8 raters blindly to each other. Accordingly, the highest agreement value was calculated in the intra-class correlation coefficient and this value was followed by Krippendorff alpha, Fleiss kappa and Cohen kappa, respectively. It has been determined that as both the number of rateds and raters decrease, the agreement values tend to decrease. On the other hand, it was determined that the difficulty level of the cards did not have a significant effect. The high generalizability coefficients showed that the test was reliably scored by the raters. When the optimization analysis was examined, it was seen that the most suitable sample size for this test was 50. Having more than 50 participants did not increase agreement.

Keywords: *Interrater Agreement, Scaling Types, Intraclass Correlation, Kappa, Krippendorff Alpha, Generalizability Theory.*

GİRİŞ

Psikoloji, eğitim bilimleri, davranış bilimleri, tıp, biyoloji vs. birçok bilim alanında elde edilen veriler sınıflara atanmakta ya da belirli bir süreklilik içerisinde puanlanmaktadır. Alan çalışanlarının puanlamalarına veya diğer bir ifade ile kararlarına göre bireyler tanı almakta, sınıflara yerleştirilmekte veya bir tema etrafında gruplandırılmaktadır; bu da araştırma ve uygulamaları önemli oranda etkilemektedir (Bıkmaz, 2011; Raykov vd., 2012). Özellikle projektif testler, açık uçlu sorular, tıbbi görüntüleme teknikleri gibi uzman değerlendirmelerine dayalı ölçme araçlarıyla puanlanan kişilerin doğru şekilde değerlendirilebilmesi için ölçümlerin puanlayıcıdan puanlayıcıya tutarlı değerler alması gerekmektedir. Aynı ölçme aracında farklı kişilerin yaptığı puan atama işlemleri üzerinden bireyler veya vakalar değerlendirilirken puanlayıcılardan / yargıcılardan kaynaklı olarak gerçekleşebilen yanlış atama işlemlerinin önüne geçmek amaçlanmaktadır. Bu bakımdan puanlayıcıların uyumlu puanlar vermeleri sınıflandırma geçerliliğinin bir parçası olarak da değerlendirilmektedir. Bu araştırmada ölçek maddelerinin sınıflayıcı, sıralayıcı ve sürekli ölçekleme düzeylerine göre seçilen yöntemlerin puanlayıcılar arası uyum düzeylerini değiştirip değiştirmediğini incelemek amaçlanmaktadır. Ayrıca puanlayıcı ve puanlanan sayıları da değişimlenerek her iki koşul için de örneklem büyüklüklerinin etkisi olup olmadığı incelenecektir.

Puanlayıcılar arası uyum (interrater agreement), diğer bir ifade ile puanlayıcılar arası güvenirlik (interrater reliability), aynı testi puanlayan iki ya da daha fazla puanlayıcı arasındaki uyum ya da tutarlılığın miktarını göstermektedir (Cohen vd., 1996; Hallgren, 2012). Ölçme ve değerlendirmenin alanına giren tüm konularda puanlayıcılar arası uyumun incelenmesi bulguların tutarlılığını arttırmaktadır. Örneğin; birden çok puanlayıcının değerlendirdiği bir sınavdan bahsedilecek olursa bu puanlayıcıların puanlamaları uyum göstermediğinde sınavı alan grubun başarı sıralaması doğru olmayacaktır (Fleiss vd., 1969). Bu sebeple puanlayıcılar arası uyum aynı zamanda puanlayıcı güvenirliği ve sınıflama güvenirliği bağlamında da ele alınmaktadır (Erkuş, 1999). Puanlayıcılar arası uyumu hesaplamak için ham uyum yüzdesi, sınıf içi korelasyon, varyans analizi (ANOVA), log-lineer analiz gibi birçok yöntem kullanılır (Bıkmaz, 2011; Bıkmaz Bilgen & Doğan, 2017). Bu araştırma kapsamında incelenen sınıf içi korelasyon, Cohen ve Fleiss'ın kappaları, Krippendoff'un alfası ve genellenebilirlik kuramı aşağıda anlatılmıştır.

Puanlayıcılar arası uyumun en klasik tekniklerinin başında Pearson korelasyon katsayısı gelmektedir. Sonraları, Pearson korelasyon temelinde sürekli değişkenler için sınıf içi korelasyon katsayısı (Intraclass Correlation Coefficient-ICC) geliştirilmiştir (Fisher, 1950). Sürekli olmayan değişkenler için Cohen ve Fleiss'ın kappa katsayılarının yanı sıra Krippendorff alfa istatistiği de kullanılmaktadır (Hayes & Krippendorff, 2007); bu yöntemler, tüm puanlayıcılar, tüm puanlama kategorilerini kullanmadığında tutarlı sonuçlar vermemekte aynı

zamanda var olan uyumun gözden kaçmasına sebep olabilmektedir (Hayes & Krippendorff, 2007; Yarnold, 2016; Zapf vd., 2016).

Puanlayıcılar arası uyumun değerlendirilmesi çalışmalarına Pearson korelasyon kullanılarak başlanmış; sonraki yıllarda ikiden fazla puanlayıcının olduğu için sınıf içi korelasyon katsayısı (SKK) geliştirilmiştir (Pearson, 1901, akt. von Eye & Mun, 2005). Pearson tarafından geliştirilen SKK Fisher (1950) tarafından formülasyonu değiştirilerek ortalama kareler yöntemi temelinde hesaplanmaya başlanmıştır. SKK, en basit anlamıyla puanlayıcı varyanslarının birbirilerine olan oranıdır ve diğer korelasyonlar gibi 0 ile 1 arasında standart bir değer alır. SKK'nin 1'e yakın değerler alması puanların birbiri ile yüksek derecede uyumlu olduğunu; 0'a yakın değerler alması ise, puanların benzer olmadığını; diğer bir ifadeyle puanlayıcılar arasında anlamlı bir uyum bulunamadığını gösterir (Koo & Li, 2010; Ateş vd., 2009).

Cohen (1960), sınıflayıcı ölçekleme için kappa (κ) katsayısını geliştirmiştir. Cohen kappa katsayısı çapraz tablolara dayanmakta ve puanlayıcıların aynı kategoride puanlama olasılığını göstermektedir. Kappa katsayısının korelasyonel yöntemlere göre en önemli üstünlüklerinden biri şans uyumunu hesaba katması ve şans düzeltmesi yapmasıdır. Elde edilen katsayı -1 ile +1 arasında değer almakta ve katsayının yüksek değerleri yüksek uyum anlamına gelmektedir. Orijinal formül 2 puanlayıcı için geliştirilmiş olsa da daha sonra çoklu puanlayıcıya uyarlanmıştır. Ancak buradan elde edilen değerler prevalans problemine sahiptir. Prevelans problemi, puanlayıcıların belli örüntüde puanlamasının; örneğin, her puanlayıcının aynı denekler olmasa dahi deneklerin %50'sine aynı puanı vermesinin, yüksek kappa değerlerine yol açması olarak tanımlanmaktadır (Hallgren, 2012). Ayrıca marjinal puanlayan puanlayıcılar uyum değerinin düşük kestirilmesine (estimation) yol açmaktadır. Landis ve Koch'a (1977) göre 0'dan küçük değerler uyumsuzluğu, 0.01-0.20 zayıf uyumu, 0.21-0.40 orta düzeyde uyumu, 0.41-0.60 kabul edilebilir uyumu, 0.61-0.80 iyi uyumu ve 0.81-1.00 ise mükemmel uyumu göstermektedir.

Daha sonra Fleiss (1971) tarafından doğrudan çoklu puanlayıcı için bir diğer kappa katsayısı geliştirilmiştir. Fleiss kappa katsayısı; puanlayıcıların, puanlayıcı evreninden seçkisiz olarak seçildiğini varsaymaktadır. Landis ve Koch'un (1977), Cohen kappa katsayısı için belirledikleri uyum düzeylerine dair kriter aralıklar Fleiss kappa için de aynıdır. Çoklu puanlayıcı için Cohen kappa değeri anlamsız çıkma eğilimindeyken Fleiss kappa değeri 2 ve daha fazla puanlayıcı için daha tutarlı sonuçlar vermektedir (Hallgren, 2012).

Krippendorff'un alfa (α) katsayısı ise sıralayıcı ölçekleme düzeyi için geliştirilmiştir ve 1970'li yıllardan itibaren değerlendirici, puanlayıcı ya da kodlayıcıların olduğu yöntemler için sıklıkla kullanılmaktadır. Bu katsayının tercih edilmesindeki en önemli sebep formülünün kayıp veriler için düzeltme

prosedürlerini içermesidir (Krippendorff, 1970, 1995, 2004a, 2004b; Zhao vd., 2018). Bu katsayı ağırlıklandırma prosedürü sayesinde sıralayıcı, sınıflayıcı ve sürekli değişken düzeylerine uyarlanabilmektedir. Bu yöntemin güçlü yanı örneklem büyüklüğü ve puanlayıcı sayısı ile ilgili düzeltmeler yapmasıdır. Temel formül gözlenen uyumsuzluk/beklenen uyumsuzluk oranına dayanmaktadır (Krippendorff, 2004a). Krippendorff alfanın 0.67'den küçük değerleri zayıf, 0.67-0.80 arası orta ve 0.80'den büyük değerleri ise yüksek uyum olarak değerlendirilmektedir (Bıkmaz Bilgen & Doğan, 2017; Hayes & Krippendorff, 2007).

Puanlayıcılar arası uyumun değerlendirilmesi amacıyla kullanılan diğer bir yöntem olan genellenebilirlik kuramı ya da G kuramı, gözlemleri karşılaştırmaya yarayan bir perspektif sunmaktadır. Genellenebilirlik kuramı Cronbach, Rajanatham ve Gleser (1963) tarafından geliştirilmiştir. Ölçümlerin uyum ya da güvenirliği incelenirken çeşitli koşullar tanımlanır. Özellikle performans değerlendirmelerinde performansa dahil olabilecek birçok değişkeni aynı anda inceler. Genellenebilirlik kuramında kontrol edilen bu değişkenler klasik yöntemlerdeki kovaryantlardan farklıdır. Kontrol değişkenlerinin düzeylerinin ya da kategorilerinin de değişimi ne yönde etkilediği tespit edilebilir. Ek olarak kontrol değişkenlerinin hangi düzeylerinin aslında ölçümü optimize ettiği de görülebilir. Klasik test kuramından farklı olarak G kuramında ölçütlerin mutlak ve göreceli olmasına göre sonuçların tutarlığı ayrı ayrı incelenebilir. Gözlenen değişkenlerin birbirlerinin kümesi olmasına izin verilir. Bu kuramda değişimin kaynakları yüzey (facet) olarak adlandırılmakta ve bu yüzeyler varyans analizindeki faktörlere benzer şekilde kullanılmaktadır. Her bir yüzey bir potansiyel hata kaynağı olarak ele alınır. Genellenebilirlik kuramı ile puanlayıcılar arası uyum incelemelerinde ölçüm objesi, puanlananların aldığı puanlar ve puanlayıcılar ayrı ayrı modellenerek her bir puanlanan, puanlayıcı ve ölçüm düzeyi için varyans oranları kestirilir. Bu kestirim sonucunda puanlayıcılar arası uyumun yüksek olması için puanlayıcıdan kaynaklanan varyansın düşük olması gerekir. Hesaplanan G katsayısı aracılığıyla en uygun uyumun kaç puanlayıcı ya da puanlanan olduğu durumda hesaplandığı da incelenebilir (Atılgan, 2019, Briesch vd,, 2014; Crocker & Algina, 1986; Shrout & Fleiss, 1979; Yıldıztekin, 2014).

Alanyazında puanlayıcılar arası uyum ile ilgili farklı katsayıların karşılaştırılmasına ve puanlayıcı ile puanlanan sayılarının değişiminin uyum düzeyi üzerindeki etkisine dayalı çalışmalar bulunmaktadır. Sınıfiçi korelasyon ve Krippendorff alfa katsayılarının karşılaştırıldığı bir çalışmada, sınıfiçi korelasyon katsayısına ait değerlerin daha yüksek olduğu görülmüştür (ten Hove, vd., 2018). Atmaz (2009)'ın 50 puanlanan ve 17 puanlayıcı ile yaptığı çalışmada, sınıfiçi korelasyon ve G katsayılarının birbiriyle uyumlu sonuçlar verdiği görülmüştür. Farklı bir çalışmada Monte Carlo simülasyonu ile üretilen veri üzerinden sınıfiçi korelasyon katsayısı ile çok puanlayıcı için düzenlenmiş Cohen kappa

hesaplanarak uyum düzeyleri varyans analizi ile karşılaştırılmış; her düzeyde Cohen kappa katsayısı, sınıfiçi korelasyondan daha düşük değerler almıştır. Aynı çalışmada puanlanan sayısı 25-200 arasında ve puanlayıcı sayısı 2-10 arasında değişimlenerek hem gerçek veri hem de simülasyon verisi incelenmiş; puanlayıcı sayısı 6 olduğunda ve puanlanan sayısı 40'ı geçtiğinde uyum düzeyinin önemli düzeyde artış göstermediği tespit edilmiştir (Nying, 2004). Abedi vd. (1995)'in yaptığı farklı bir simülasyon çalışmasında örneklem büyüklüğü arttıkça kappa, alfa ve SKK düzeylerinin arttığı belirlenmiş; ancak bu artış puanlanan sayısının 30 olduğu durumdan itibaren düzleşme eğilimi göstermiştir. Öte yandan bu çalışmada puanlayıcı sayısının 4'ten fazla olduğu koşullarda, uyum değerlerinin farklılaşmadığı görülmüştür.

Bu araştırmanın temel amacı farklı ölçekleme düzeylerine göre ele alınan yöntemlerde farklı puanlanan ve puanlayıcı sayıları için puanlayıcılar arası uyumun farklılaşıp farklılaşmadığını incelemektir. Bu doğrultuda araştırmanın temel soruları şunlardır:

1- SKK, Cohen kappa, Fleiss kappa ve Krippendorff alfa uyum katsayıları 3 farklı puanlayıcı sayısı ve 3 farklı puanlanan sayısına göre farklılaşmakta mıdır?

2- Puanlayıcı sayısı, puanlanan sayısı, uyum yöntemleri ve test maddeleri olan figür/kart düzeylerine göre varyans kaynakları nelerdir?

3- Genellenebilirlik kuramına göre optimal puanlanan ve puanlayıcı sayısı kaçtır?

YÖNTEM

Çalışma Grubu ve Katılımcılar

Araştırmanın puanlananlarını Korkmaz ve diğerleri (2019, 2022) tarafından yürütülen Bilimsel Araştırma Projesi kapsamındaki Bender Gestalt II (BGT-II) testinin Kopyalama figürlerini yanıtlayan 4-10 yaş aralığındaki (ortalama yaş 4.89, S=1.66) 35 çocuk oluşturmaktadır. Puanlananların %48.6'sı kız (n=17) ve %51.4'ü erkek (n=18) çocuktur. Araştırmada 35 katılımcı çocuk, 8 farklı puanlayıcı tarafından değerlendirilmiştir. Puanlayıcılar psikoloji lisans eğitimini tamamlamış 6 kadın ve 2 erkek psikologdan oluşmaktadır.

Veri Toplama Aracı

Bender Gestalt II Kopyalama maddeleri (Brannigan & Decker, 2003) veri toplama aracını oluşturmuştur. Bender Gestalt II görsel motor bütünleştirme yeteneğini ölçmek üzere toplamda 16 farklı karttan oluşmaktadır. Her kartta farklı seviyelerden figürler bulunmaktadır ve katılımcılar tarafından çizilen figürlerin

orijinal figüre benzerliği temelinde Likert formatında 0-4 (0-benzerlik yok ve 4: mükemmel benzerlik) aralığında puanlanmaktadır. Ayrıca bu testte kullanılan kartlar ilerledikçe kopyalama görevinin zorlaştığı kabul edilmektedir. Bu nedenle ilk 3 kart 7 yaşından büyük katılımcılara, son 4 kartı ise 7 yaşından küçük katılımcılara uygulanmamaktadır. Araştırmada 35 katılımcının (puanlanan) birbirine kör 8 puanlayıcı (6 kadın, 2 erkek) tarafından puanlandığı Kopyalama Testine ait 16 karttan tüm yaş grupları için ortak kullanılan 9 kart veri olarak kullanılmıştır.

İşlem

Araştırmada, 3 farklı puanlayıcı düzeyi (4, 6 ve 8 puanlayıcı) ve 3 farklı katılımcı sayısı düzeyi (10, 20 ve 35 katılımcı) üzerinden değerlendiriciler arası uyum değerleri hesaplanmıştır. Tüm koşullar için sınıf içi korelasyon katsayısı, Cohen'in kappa, Fleiss'ın kappa değeri ve son olarak Krippendorff'un alfa katsayısı hesaplanmıştır.

Genellenebilirlik kuramı kapsamında Bender-Gestalt II Kopyalama maddeleri/kartları, puanlanan sayısı, puanlayıcı sayısı ve uyum yöntemleri yüzey olarak tanımlanarak varyans kaynakları tespit edilmiş ve optimizasyon çalışması yapılmıştır.

Verilerin Analizi

Sınıf içi korelasyon katsayısı ve Fleiss kappa değerleri doğrudan IBM SPSS 25.0 paket programı kullanılarak hesaplanmıştır. Cohen kappa değeri için Cambridge Üniversitesi sayfasından alınan David Nichols tarafından 1997'de yazılmış ve 2013'te güncellenmiş SPSS sentaksı kullanılmıştır. Krippendorff'un alfa değerleri için ise Hayes ve Krippendorff (2007) tarafından geliştirilen Kalpha 4.0 makrosundan yararlanılmıştır. Yöntemler arasındaki farklılaşmaların istatistiksel anlamlılığını incelemek amacıyla genel doğrusal modeller kullanılmış ve analizler IBM SPSS 25.0 paket programında yapılmıştır. G katsayıları ve optimizasyon çalışması için hesaplamalar EduG paket programı kullanılarak yapılmıştır (Cardinet vd., 2009).

BULGULAR

Puanlayıcılar Arası Uyum Katsayılarının İncelenmesi

Öncelikle tüm puanlayıcıların (8 değerlendirici) modelde bulunduğu durumda 35 katılımcının tamamı için katsayılar incelenmiştir. Sınıf içi korelasyona göre en yüksek uyum değerleri r=0.73 (p<0.001) ile 5 ve 12. kartlarda, Cohen kappaya göre κ=0.32 (p<0.001) ile 12. kartta, Fleiss kappaya göre κ=0.38 (p<0.001) ile 12. kartta ve son olarak Krippendorff alfaya göre ise α=0.69 (p<0.001) ile 12. kartta hesaplamıştır. Bu koşul için tüm katsayılar incelendiğinde her durumda 12. kartın daha uyumlu puanlandığı görülmüştür. Daha sonra

katılımcı sayısı değiştirilmeden puanlayıcı sayısı 6'ya düşürülmüştür. Bu durumda da 12. kartın her durumda en uyumlu puanlanan kart olduğu görülmüştür, r=0.72 (p<0.001), *Cohen* κ=0.35 (p<0.001), *Fleiss* κ=41 (p<0.001), Krippendorff α=0.68 (p<0.001). Katılımcı sayısının 35 olduğu durum için son olarak 4 puanlayıcının puanlamaları üzerinden hesaplamalar yapılmıştır. Bu durumda en yüksek sınıf içi korelasyon değerine 10. kartta (r=0.81) ulaşılmakla beraber diğer katsayılar kullanıldığında en yüksek uyumu 12. kartın göstermeye devam ettiği (*Cohen* κ=0.21, *Fleiss* κ=0.31, Krippendorff α=0.66) görülmüştür, tüm katsayılar için p<0.001'dir. Genel olarak 35 katılımcının olduğu her durumda en yüksek değerleri sınıf içi korelasyon katsayısının verdiği tespit edilmiştir. Sınıf içi korelasyonu sırasıyla Krippendorff α, Fleiss κ ve en düşük değerle de Cohen κ takip etmektedir. Bu noktada ölçümün sürekli kabul edildiği durumlarda en yüksek, sınıflayıcı kabul edildiği durumlarda ise en düşük uyumun hesaplandığı gözlenmiştir, bkz. Tablo 1.

Daha sonra puanlanan sayısı 20'ye düşürülmüş ve bu koşul için analizlere başlanmıştır. İlk olarak 8 puanlayıcı desene dahil edilmiştir. Bu durumda da hem sınıf içi korelasyon katsayısına (r=0.79, p<0.001) hem Cohen (κ=0.36, p<0.001) ve Fleiss'ın (κ=0.47, p<0.001) kappa katsayılarına hem de Krippendorff'un alfa (α=0.78, p<0.001) katsayısına göre 12. kartın daha uyumlu puanlandığı tespit edilmiştir. Devamında 6 puanlayıcı için analiz yapılmış ve bu analizler sonucunda da tüm katsayılar incelendiğinde yine en yüksek uyumun 12. kartta gösterildiği (r=0.71, *Cohen* κ=0.21, *Fleiss* κ=0.31, α=0.66) tespit edilmiştir, tüm katsayılar için p<0.001. Puanlayıcı sayısı 4'e düşürüldüğünde sınıf içi korelasyon (r=0.77, p<0.001), Fleiss kappa (κ=0.35, p<0.001) ve Krippendorff alfa (α=0.72, p<0.001) katsayılarına göre en uyumlu kart 12 olmaya devam ederken Cohen kappa (κ=0.18, p<0.05) katsayısına göre 5. kartın öne çıktığı tespit edilmiştir.

Son olarak puanlanan sayısı 10'a düşürülmüştür. Puanlayıcı sayısının 8 olduğu tüm durumlarda 12. kartın en uyumlu kart olduğu tespit edilmiştir. Ancak 6 puanlayıcının olduğu koşul incelendiğinde sınıf içi korelasyon katsayısının (r=0.83, p<0.001) ve Krippendorff'un alfasının (α=0.68, p<0.001) en yüksek değeri 10. kartta hesaplanmıştır. Diğer taraftan Cohen kappa (κ=0.32, p<0.001) ve Fleiss kappa (κ=0.32, p<0.001) katsayılarına göre 12. kart yerini koruyarak en uyumlu kart olmaya devam etmiştir. Puanlayıcı sayısının 4 olduğu son koşula gelindiğinde sınıf içi korelasyon katsayına (r=0.66, p<0.001) göre 13. kart, Cohen kappa katsayısına (κ=0.23, p<0.05) göre 12. kart, Fleiss kappa katsayısına (κ=0.44, p<0.001) göre 10. kart ve son olarak Krippendorff alfa katsayısına (α=0.74, p<0.001) göre 10. kartta en yüksek uyum değeri hesaplanmıştır.

Genel olarak tüm koşullar göz önüne alındığında bu çalışmanın ele alınan her koşulunda en yüksek uyumun puanlamanın sürekli kabul edildiği durumlarda kullanılan sınıf içi korelasyon yönteminde hesaplandığı görülmüştür. Ardından Krippendorff'un alfası, Fleiss'ın kappası ve Cohen'in kappası gelmiştir. Puanlanan ve puanlayıcı sayısına göre sınıf içi korelasyon, Cohen'in kappa katsayısı, Fleis'ın

kappa katsayısı ve Krippendorff'un alfasına ait medyan değerleri Tablo 1'de gösterilmiştir.

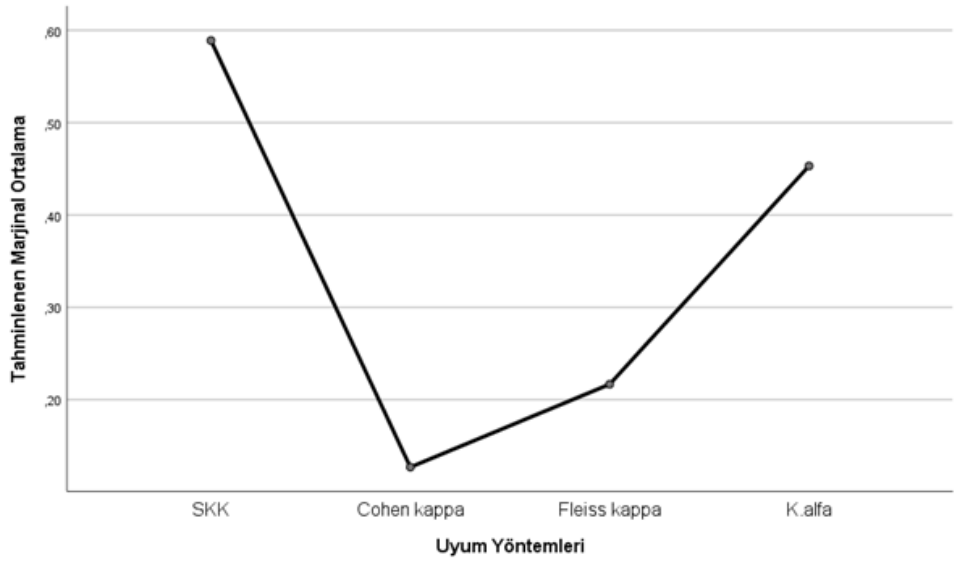
Tablo 1: *Farklı Puanlayıcı ve Puanlanan Sayısına Göre Uyum Katsayılarının Medyan Değerleri*

Puanlanan Sayısı	Puanlayıcı Sayısı	Sınıf İçi Korelasyon	Cohen Kappa	Fleiss Kappa	Krippendorff Alfa
	8	0.66	0.17	0.25	0.54
35	6	0.61	0.17	0.24	0.51
	4	0.67	0.10	0.21	0.54
	8	0.61	0.18	0.25	0.52
20	6	0.56	0.17	0.24	0.51
	4	0.66	0.10	0.21	0.48
	8	0.56	0.12	0.15	0.41
10	6	0.55	0.12	0.17	0.37
	4	0.62	0.04	0.11	0.33

Varyans Analizi Bulguları

Uyum yöntemlerinden elde edilen değerlerin ortalama düzeyinde farklılaşmasını incelemek amacıyla tekrarlı ölçümler varyans analizi yapılmıştır. Mauchly'nin küresellik testi kullanılmış ve küresellik varsayımı karşılanamadığından (p<0.001) Huyn-Feldt düzeltmesi uygulanmıştır. Analiz sonuçları incelendiğinde farklı yöntemler kullanılarak hesaplanan uyum katsayılarının istatistiksel olarak anlamlı derecede farklılaştığı saptanmıştır, F(1.38, 109.23)=323.90, η^2=0.80, *güç*=1.00, p<0.001. En düşük uyum değerinin Cohen κ ($\bar{x}$=0.13, σ=0.13) olduğu ve bu katsayıyı Fleiss κ ($\bar{x}$ =0.22, σ=0.11) ile Krippendorff α'nın ($\bar{x}$=0.45, σ=0.19) izlediği görülmüştür. En yüksek uyum değeri ise ölçümün sürekli olduğu varsayımını barındıran sınıf içi korelasyon yöntemi ($\bar{x}$=0.59, σ=0.18) ile hesaplanmıştır, bkz. Şekil 1.

Şekil 1: Farklı Uyum Yöntemlerinden Elde Edilen Ortalamaların Karşılaştırılması



Puanlayıcı ve puanlanan sayısına göre uyum değerlerinin anlamlı düzeyde farklılaşıp farklılaşmadığını görmek amacıyla elde edilen veriye çok değişkenli varyans analizi (MANOVA) uygulanmıştır. Box M testine göre varyans homojenliği varsayımının karşılandığı görülmüştür, $F(80, 6050.19)=1.06$, $M=108.25$, $p=0.33$. Model anlamlılığının Pillai izi kriterine göre sağlandığı saptanmıştır, $F(4, 68)=318.51$, $\eta^2=0.95$, *güç*=1.00, $p<0.001$. Pillai izi, varyans analizlerinde varsayımlardan sapma olduğunda ya da örneklem büyüklüğü küçük olduğunda kullanılması önerilen anlamlılık kriteridir (Meyers vd., 2013; Pillai, 1955; Tabachnick & Fidel, 2013). Araştırmada puanlanan sayısının az olması sebebiyle bu yöntemin kullanılması uygun görülmüştür. Ancak hem puanlanan ($p=0.11$) hem de puanlayıcı ($p=0.18$) değişkenlerinin anlamlılığı Pillai izi kriterine göre incelendiğinde anlamlı farklılaşma bulunamamıştır.

Genellenebilirlik Analizi Bulguları

Puanlayıcılar arası uyum için saptanan uyum katsayıları arasındaki farklardaki varyans kaynaklarının açıkladığı alanları incelemek amacıyla; uyum katsayılarının kart düzeyi, puanlanan sayısı ve puanlayıcı sayısına göre değişiminin modellendiği (Yöntem/Kart*Puanlayıcı*Puanlanan) genellenebilirlik analizi uygulanmıştır. Öte yandan bu analiz dahilinde puanlayıcı ve puanlanan sayısının artması durumunda genellenebilirlik düzeyinin nasıl değişeceğini görmek amacıyla 50, 100, 500 ve 1000 kişilik puanlayıcı ve puanlanan koşulları için optimizasyon hesaplaması yapılmıştır.

Analiz sonuçlarında uyumdaki değişimin varyansının %52.5'inin uyum yöntemleri tarafından açıklandığı görülmüştür. Toplam varyansın %16'sı kartlar tarafından, %1.1'i puanlayıcılar tarafından ve %3.8'i ise puanlananlar tarafından açıklanmaktadır. Ayrıca kart*puanlanan (%4.1) ve kart*yöntem (%14.1) kesişimleri de varyansı açıklamaktadır. Genellenebilirlik kuramına göre en yüksek değişim kaynaklarının kullanılan yöntemler ve kart düzeyleri olduğu görülmektedir, bkz, Tablo 2.

Tablo 2: *Uyum Değerlerinin Değişimindeki Varyans Kaynakları*

	Toplam Kareler	Ortalama Kareler	Standart Hata	Açıklanan % Varyans
Kart	3.261	0.408	0.005	16.2
Puanlayıcı	0.153	0.077	0.001	1.1
Puanlanan	0.550	0.275	0.002	3.8
Yöntem	11.010	3.670	0.029	52.5
Kart*Puanlayıcı	0.146	0.009	0.000	0.5
Kart*Puanlanan	0.554	0.035	0.001	4.1
Kart*Yöntem	2.095	0.087	0.003	14.1
Diğer (Toplam)				7.4 (100)

Bu çalışmaya ait göreceli (relative) genellenebilirlik katsayısı 0.98 ve mutlak (absolute) genellenebilirlik katsayısı 0.94 olarak belirlenmiştir. Genellenebilir yüzeyler (G-Facets) analizine göre kart düzeyinde en yüksek göreceli G katsayısı 0.986 ile 2. düzey (Kart 6) kartta, puanlayıcı düzeyinde 0.982 ile 6 puanlayıcının olduğu durumda ve son olarak puanlanan sayısı bakımından da 0.983 ile 10 katılımcının olduğu durumda tespit edilmiştir. Mutlak ve göreceli G katsayıları Tablo 3'te gösterilmiştir.

Tablo 3: *Genellenebilir Yüzeyler Analizi Bulguları*

Yüzey	Düzey	Göreceli G Katsayısı	Mutlak G Katsayısı
Kart	5 (Düzey 1)	0.974	0.920
	6 (Düzey 2)	0.986	0.956
	7 (Düzey 3)	0.975	0.925
	8 (Düzey 4)	0.975	0.922
	9 (Düzey 5)	0.975	0.930
	10 (Düzey 6)	0.977	0.908
	11 (Düzey 7)	0.979	0.943
	12 (Düzey 8)	0.977	0.938

	13 (Düzey 9)	0.980	0.938
Puanlayıcı	8	0.977	0.930
	6	0.982	0.936
	4	0.980	0.939
Puanlanan	35	0.980	0.908
	20	0.980	0.914
	10	0.983	0.963

Son olarak optimizasyon çalışması sonuçları incelenmiştir. Bu çalışmaya göre 50 puanlayıcı ve 50 puanlanan olduğu durum için göreceli G katsayısı 0.986 (*mutlak G*=0.971), 100 puanlayıcı ve 100 puanlanan olduğu durum için göreceli G katsayısı 0.986 (*mutlak G*=0.971), 500 puanlayıcı ve 500 puanlanan olduğu durum için göreceli G katsayısı 0.986 (*mutlak G*=0.971), son olarak 1000 puanlayıcı ve 1000 puanlanan olduğu durum için göreceli G katsayısı 0.986 (*mutlak G*=0.971) olarak hesaplanmıştır. Puanlayıcı sayısı 8'den 50'ye ve puanlanan sayısı 35'ten 50'ye çıktığında göreceli uyum katsayısı 0.982'den 0.986'ya az da olsa yükselmesine rağmen örneklemin 50'den sonraki artışlarında 0.986 olarak kalmaya devam etmiştir. Mutlak G katsayıları incelendiğinde ise eldeki veriye ait 0.940 olan değer 50 puanlayıcı ve puanlanan olduğu durumda 0.971'e yükselirken puanlayıcı ve puanlanan sayısının daha çok artmasının bu noktadan sonra bir değişime yol açmadığı tespit edilmiştir.

SONUÇ ve TARTIŞMA

Araştırmada ilk olarak klasik yöntemlere göre puanlayıcılar arası uyum incelenmiştir. Tüm koşullar göz önüne alındığında her zaman en yüksek uyumu sürekli ölçekleme düzeyi için kullanılan sınıf içi korelasyon yöntemi sağlamıştır. Sınıf içi korelasyon yöntemini Krippendorff'un sıralayıcı ölçekleme düzeyi için formüle edilmiş alfa katsayısı takip etmiştir. Cohen kappa temel formülü itibariyle iki puanlayıcının olduğu duruma göre tasarlanmış olsa da daha sonra değiştirilen formülle çoklu puanlayıcı durumuna genelleştirilmiştir. Fleiss kappa formülü ise doğrudan çoklu puanlayıcı için geliştirilmiş olsa da Cohen kappa temelinde ve Cohen kappanın bir türevi şeklindedir. Bu iki kappa katsayısının tüm analizlerde birbirine çok yakın sonuçlar verdiği ancak Fleiss kappanın genel olarak daha yüksek olduğu gözlenmiştir. En düşük uyum değerleri sınıflayıcı ölçekleme için kullanılan kappa formülleri ile tespit edilmiştir. Puanlanan sayısının azalmasının değerlerin anlamsız çıkma eğilimine yol açtığı görülmüştür. Aynı şekilde puanlayıcı sayısı azaldıkça da hesaplanan katsayı değerleri azalmıştır. Bu bulgulardan elde edilen değerler için varyans analizleri yapılmış ve elde edilen uyum düzeylerinin farklılaştığı istatistiksel olarak ortaya konmuştur. Fleiss kappanın Cohen'in katsayısına göre uyumu daha iyi açıkladığı birçok araştırmada

gösterilmiştir (Bıkmaz, 2011; Fleiss, 1971; Landis & Koch, 1977). Öte yandan puanlayıcılar arası uyumun değişken yapısının sürekli mi süreksiz mi olduğuna göre değiştiği de ortaya konmuştur (Bıkmaz, 2011; Hayes & Krippendorff, 2007). Bu çalışmada, kullanılan uyum yöntemine göre elde edilen bulguların birbirinden farklılaştığı ortaya konmuştur. Bu da kullanılan ölçüm aracının ölçekleme düzeyine doğru karar verilmesinin önemini ispatlamaktadır. Örneğin sınıf içi korelasyon kullanıldığında orta-yüksek bir uyumdan bahsedilebilecekken kappa katsayısı kullanıldığında elde edilen puanların uyumsuz olduğu söylenebilmektedir.

Klasik yöntemlerden elde edilen değerler incelendiğinde kullanılan kartların zorluk düzeylerinin ise genel bir örüntü sergilemediği, zorlaşmanın uyumu arttırıcı ya da azaltıcı bir etkisinin olmadığı görülmüştür. Bunun istatistiksel olarak sınanması açısından varyans analizleri sırasında "zorluk düzeyi" analize kovaryant olarak eklenmiş ve kart düzeyinin uyum düzeyi üzerinde anlamlı bir etkisi olmadığı görülmüştür. Bunun aksine genellenebilirlik kuramına göre elde edilen bulgularda kart düzeyinin en büyük değişkenlik kaynaklarından biri olduğu tespit edilmiştir. Klasik kuramlar ve genellenebilirlik kuramının farklı sonuç vermesi literatürde de karşılaşılan bir durumdur (Atılgan, 2019; Yıldıztekin, 2014).

Bu çalışmanın sonuçlarına göre puanlanan ve puanlayıcı sayısının değişiminin uyum üzerinde anlamlı bir etkisinin olmadığı görülmüştür. Bu durum literatürle çelişmektedir. Hem puanlanan hem de puanlayıcı sayısının artmasının özellikle bu çalışmanın konusu olan klasik uyum yöntemleri için uyum düzeyini arttırıcı bir etkisi olduğu bilinmektedir (Bıkmaz, 2011; Hayes & Krippendorff, 2007; von Eye & Mun, 2005). Ancak yine birçok çalışmada da puanlayıcı ve puanlanan düzeyi artışının belli noktalardan sonra çok bir şey değiştirmediği görülmüştür. Buradaki optimizasyon çalışmasında olduğu gibi sayı arttıkça bir noktaya kadar klasik uyum yöntemlerinin büyüklüğü artsa da genellikle bir süre sonra doygunluğa ulaşmaktadır (Abedi vd. ,1995). Örneğin Arslan Mancar'ın 2019 tarihli çalışmasında 2 puanlayıcının uyumunun 3 ve 5 puanlayıcıdan daha düşük olduğu ancak 3 ile 5'in genellikle yakın sonuçlar verdiği görülmüştür. Ayrıca Saito vd. (2006) yaptıkları çalışmada puanlanan sayısı sabit tutulduğunda puanlayıcı sayısı arttırıldığında uyumun 4 puanlayıcıda doygunluğa ulaştığını bu noktadan sonraki artışlar arasında anlamlı fark olmadığını bulmuşlardır. Yine aynı çalışmada puanlayıcılar arası varyans artışı yoksa daha fazla puanlayıcıya ihtiyaç duyulmayacağı belirtilmiştir. Dolayısıyla bu çalışmanın bulguları, her bir puanlayıcının her bir katılımcıyı değerlendirdiği, puanlama ölçütlerinin önceden belirlendiği ve bir puanlama anahtarının olduğu (Bender Görsel Motor Gestalt testi gibi) sürekli değişken formatındaki test/ölçek/soru maddelerinin farklı sayıdaki puanlayıcılar arası uyum düzeylerinin tutarlığının genel anlamda puanlanan sayısı ve yöntemleri ile sınırlıdır.

Araştırma bulguları genellenebilirlik kuramı ile incelendiğinde puanlayıcılar arası uyum değişiminin %52.5'i uyum yöntemleri tarafından açıklanmıştır. Dolayısıyla varyans kaynakları olarak değişim etkisinin en temel belirleyicisinin incelenen yöntemler olduğu tespit edilmiştir. Ayrıca puanlayıcılar tarafından açıklanan varyansın %1.1 olduğu; düşük değerdeki bu değişimin puanlayıcılar arası tutarlılık ve uyumu gösterdiği belirlenmiştir. Hem göreceli genellenebilirlik katsayısı 0.98 hem de mutlak genellenebilirlik katsayısı 0.94 derecelerinde saptanmış olup araştırma deseninin güvenilir olduğu görülmüştür. Puanlayıcı ve puanlananın en uygun sayısının belirlenmesi amacıyla yapılan optimizasyon incelemesi sonucunda her iki koşul için de bu değerin 50 olduğu saptanmıştır. Puanlanan ve puanlayıcı sayısının 50 gibi bir değer olması bazı açılardan tartışmaya açıktır. Örneğin yapılandırılmış bir ölçme aracı üzerinden gözlem yapılarak değerlendirilmesi veya söz konusu yapılandırılmış ölçme aracının standart bir puanlama anahtarına göre uzmanlar tarafından puanlanması gibi durumlarda değerlendirici sayısının genellikle daha az olduğu bilinmektedir. Nitekim çalışma bulguları incelendiğinde 8 puanlayıcı için G katsayısının 0.98 olduğu, puanlayıcı sayısı 50 olduğunda ise 0.99'a yükseldiği görülmektedir. Benzer şekilde 35 puanlanan için G katsayısı 0.98 iken puanlanan sayısı 50 olduğunda bu değer 0.99'a yükselmektedir. Sonuç olarak puanlayıcı ve puanlanan sayısındaki artışın genellenebilirlik düzeyleri üzerindeki etkisinin küçük olduğu ve 8 puanlayıcı-35 puanlanan ile de son derece yüksek ve uyumlu sonuçlara ulaşıldığı belirlenmiştir. Dolayısıyla puanlayıcı sayısının optimizasyonla önerilen 50 gibi bir sayı olmasının pratikte işlevsel ve ekonomik bir yönü de bulunmamaktadır. Daha önceki çalışmalarda da puanlayıcı ve puanlanan sayısı artışının bir noktadan sonra önemli bir değişime sebep olmadığı ortaya konmuştur (Abedi vd., 1995; Atmaz, 2009; ten Hove vd., 2018)

Araştırmanın vurgulanmak istenen en önemli kısım ölçekleme düzeylerine göre verilecek kararların nasıl birbirinden farklılaştığıdır. Psikometrik diğer ölçümler de düşünüldüğünde, ölçekleme psikometrinin belki de en önemli konularından biridir. Likert ve benzeri ölçeklerin kullanıldığı durumlarda da verilerin sürekli olduğu varsayımıyla hareket edilmektedir. Ancak hem Likert tipi ölçeklerde hem buradaki gibi sunî bir objektif puanlama anahtarı bulunan ölçümlerde hem de dolaylı ölçümlerde puanların sürekli olup olmadığı ciddi bir tartışma alanıdır. Daha önce de birçok araştırmacı bu konuya değinmiştir. Örneğin Likert tarafından 1932'de yayınlanan kitapçıkta belli başlı tartışma konularından biri budur. Ölçeklemenin en avantajlı tarafları; anket tipi çalışmalar için evrensel, kolay anlaşılabilir ve kolay uygulanabilir bir yöntem olmasıdır. Cevaplar ve puanlamalar bazı matematiksel anahtarların hesaplanmasına olanak tanısa da subjektiftir. Fakat psikolojik özellikler (tutum, davranış vb.) çok boyutlu bir süreklilik barındırmasına rağmen ölçeklemeler tüm boyutları temsil etmemektedir ve 1-5 (veya 1-3, 1-7, 1-11 vb.) arasında bir seçim şansı sunarlar. Öte yandan 1 ve

5 arasındaki 4 aralığın gerçekten eşit olmadığı da bir sorun olarak durmaktadır (Likert, 1932).

Ölçümlerin genellikle sürekli kabul edilmesinin bir sebebi de parametrik işlem yapmaya imkân tanımasıdır. Parametrik işlemler genellikle parametrik olmayanlara göre daha anlamlı ve güçlü olarak algılanmaktadır. Parametrik istatistiklerin karmaşıklığı da bu algıya yol açmaktadır. Araştırmacıların amaçlarından biri alanı ilerletmek için geçerli istatistikler üretmektir. Parametrik istatistiklerin anlamlı çıkma ihtimalinin daha yüksek olması da bu amacı destekler (Bishop & Herron, 2015). Birçok psikometrist bu tip ölçeklerin aslında sürekli olmadığını, sıralayıcı olduğunu belirtmektedir. Ancak öte yandan birçok araştırmacı da psikolojik ölçümlerde kullanılan ölçeklerin eşit aralıklı olduğu yönünde fikir bildirmektedir. Pearson korelasyon veya varyans analizi gibi yöntemler kullanıldığında tek maddelik Likert gibi yapay aralıklı ölçümleri kullanmak uygun değildir. Parametrik olmayan analizlerin parametrikler kadar değer görmemesi önemli bir önyargıdır ve bilimin güvenirlik ve geçerliğini etkilemektedir (Carifio & Perla, 2008; Doğan & Doğan, 2014).

Etik Kurul Beyanı: Araştırmada analiz aşamasında kullanılan veriler**,** Korkmaz vd.,(2015-2019) tarafından yürütülen *Bender Gestalt-II Testinin Koppitz II ve Bender Gestalt II Puanlama Sistemlerine Göre (4-7 ve 8-18 Yaş) Ön Norm Çalışması* (Ege Üniversitesi Bilimsel Araştırma Projesi, Proje No: 15-EDB-021) kapsamında, İzmir İl Milli Eğitim Müdürlüğünden Etik Kurul onayına sahiptir (30/12/2015 tarihli 12018877-604.01.02-E.13519131).

Yazar Katkıları ve Çıkar Çatışması: Yazarların katkı oranı eşittir ve çıkar çatışması bulunmamaktadır.

KAYNAKÇA

Abedi, J., Baker, E. L. & Herl, H. (1995). Comparing reliability indices obtained by different approaches for performance assessments. Los Angeles: University of California, CSE Technical Report, 401.

Arslan Mancar, S. (2019). *Performansa dayalı durum belirlemede puanlayıcılar arası güvenirlik tekniklerinin karşılaştırılması*. Yayınlanmış yüksek lisans tezi, Ankara Üniversitesi, Ankara.

Ateş, C., Öztuna, D. & Gen. Y. (2009). Sağlık araştırmalarında sınıf içi korelasyon katsayısının kullanımı. *Türkiye Klinikleri 1* (2), 59-64.

Atılgan, H. (2019). *Genellenebilirlik Kuramı ve Uygulaması* (1. Basım). Ankara: Anı Yayınları.

Atmaz, G. (2009). *Puanlama yönergesi kullanılması durumunda puanlayıcı güvenirliğinin incelenmesi*. Yayınlanmamış yüksek lisans tezi, Mersin Üniversitesi, Mersin.

Bıkmaz, Ö. (2011). *Üst düzey zihinsel özelliklerin ölçülmesinde puanlayıcılar arası güvenirlik belirleme tekniklerinin karşılaştırılması*. Yüksek lisans tezi. Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü Eğitim Bilimleri Ana Bilim Dalı Eğitimde Ölçme ve Değerlendirme Bilim Dalı, Ankara.

Bıkmaz Bilgen, Ö. & Doğan, N. (2017). Puanlayıcılar arası güvenirlik belirleme tekniklerinin karşılaştırılması. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *8* (1), 63-78.

Bishop, P. A. & Herron R. L. (2015). Use and Misuse of the Likert Item Responses and Other Ordinal Measures. *International Journal of Exercise Science*, *8* (3), 297-302.

Brannigan, G. G. & Decker, S. L. (2003). *Bender Gestalt II Bender Visual-Motor Gestalt Test* (Second Edition). Itasca, IL: Riverside Publishing.

Briesch, A. M., Swaminathan, H., Welsh, M. & Chafouleas, S. M. (2014). Generalizability theory: a practical guide to study design, implementation, and interpretation. *Journal of School Psychology, 52* (1), 13-35.

Cardinet, J., Johnson, S. & Pini, G. (2009). *Applying Generalizability Theory using EduG*. New York: Routledge Academic.

Carifio, J. & Perla, R. (2008). Resolving the 50-year debate around using and misusing Likert scales. *Medical Education*, *42* (12), 1150–1152.

Cohen (1960). A coefficient of rater agreement for nominal scales. *Educational and Psychological Measurement*, *20* (1), 37-46.

Cohen, J. R., Swerdlik, E. M. & Phillips, S. M. (1996). *Psychological Testing and Assessment* (Third Edition). London: Mayfield Publishing Compony.

Crocker, L., & Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. New York: Harcourt Brace.

Cronbach, L. J., Rajanatham, N. & Gleser, G. C. (1963). Theory of generalizability: a liberation of reliability theory. *The British Journal of Statistical Psychology, 16* (2), 137-163.

Doğan, İ & Doğan, N. (2014). *Adım Adım Çözümlü Parametrik Olmayan İstatistiksel Yöntemler,* 4-5. Ankara: Detay Yayıncılık.

Erkuş, A. (1999). *Ölçme araçlarının tutarlı ölçme ve sınıflama yapıp yapmadığını belirlemeye yönelik bir araştırma*. Doktora Tezi. Ankara Üniversitesi Sosyal Bilimler Enstitüsü Eğitimde Psikolojik Hizmetler Ana Bilim Dalı, Ankara.

Fisher, R. A. (1950). *Statistical Method for Research Workers* (Eleventh Edition). Edinburg: Oliver and Boyd.

Fleiss, J. L., Cohen, J. & Everitt, B. S. (1969). Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin, 72* (5), 323–327.

Fleiss, J. L. (1971). Measuring agreement for multinomial data. *Psychological Bulletin, 76* (5), 378-382.

Hallgren, K. (2012). Computimg inter-rater reliability for observational data: an overview and tutorial. *Tutorials in Quantitative Methods for Psychology, 8* (1), 23-34.

Hayes, A. F. & Krippendorff, K. (2007). Answering the call for a standard reliability measuse for coding. *Communication Methods and Measures, 1* (1), 77-89.

Korkmaz, M., Demiral, N., Sapmaz-Yurtsever, S., Kaçar-Başaran, S. & Çabuk, T.(2019). *Bender Gestalt-II Testinin Koppitz II ve Bender Gestalt II Puanlama Sistemlerine Göre (4-7 ve 8-18 Yaş) Ön Norm Çalışması*, Ege Üniversitesi Bilimsel Araştırma Projesi, Proje No: 15-EDB-021, İzmir.

Korkmaz, M., Sapmaz-Yurtsever, S.,Kaçar-Başaran,S., Demiral, N. & Çabuk, T. (2022). Bender-Gestalt II Test: Psychometric Properties with Global Scoring System on a Turkish Standardization Sample, *Child Neuropsychology*, https://doi.org/10.1080/09297049.2022.2104237.

Koo, T. K. & Li, M. Y. (2010). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15* (2), 155-163.

Krippendorff, K. (1970). Estimating the reliability, systematic error, and random error of interval data. *Educational and Psychological Measurement, 30* (1), 61–70.

Krippendoff, K. (1995). On the reliability of unitizing continious data. *Sociological Methodology, 25*, 47-76.

Krippendorff, K. (2004a). *Content Analysis An Introduction to Its Methodology* (Second Edition). Thousand Oaks, CA: Sage Publication.

Krippendorff, K. (2004b). Reliability in content analysis some common misconceptions and recommendations. *Human Communication Research, 30* (3), 411-33.

Landis, J. R. & Koch, G. G. (1977) An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics, 33* (2), 363-374.

Likert, R. (1932). *A Technique for Measurement of Attitudes* (First Edition). New York University Archives of Psychology : New York.

Meyers, L. S., Glenn G. & Guarino, A. J. (2013). *Applied Multivariate Research Design and Interpretation* (Second Edition). California: Sage.

Nichols, D. (8 Mart 2013). FAQ / kappa /multiple. 7 Şubat 2018, https://imaging.mrc-cbu.cam.ac.uk/statswiki/FAQ/kappa/multiple.

Nying, E. (2004*). A comparative study of interrater reliability coefficients obtained from different statistical procedures using monte carlo simulation tecniques.* Doctoral Dissertation. Available from Proquest Dissertations and Theses database. (UMI No. 3138768).

Pillai, K. C. S. (1955). Some new test criteria in multivariate analysis. *The Annals of Mathematical Statistics, 26* (1), 117–21.

Raykov, T., Dimitrov, D. M., von Eye, A. & Marcoulides, G. A. (2012). Interater Agreement Evaluation: a latent variable modeling approach. *Educational and Psychological Measurement, 20* (10). 1-20.

Saito, Y., Sozu, T., Hamada, C. & Yoshimura, I. (2006). Effective number of subjects and number of raters for inter-rater reliability studies, *Statistcs in Medicine, 25*, 1547-1560.

Shrout, P. E. & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin, 86* (2), 420–428.

Tabachnick, B. G. & Fidel, L. S. (2013). *Using Multivariate Statistics* (Sixth Edition). Pearson: New Jersey.

ten Hove, D., Jorgensen, T. D., & van der Ark, L. A. (2018). On the usefulness of interrater reliability coefficients. Quantitative Psychology: *The 82nd Annual Meeting of the Psychometric Society, Zurich, Switzerland*, 67-75.

von Eye, A. & Mun, E. Y. (2005). *Analyzing rater agreement manifest variable methods* (First Edition). Mahwah, New Jersey London: Lawrence Erlbaum Associates.

Yarnold, P. R. (2016). ODA vs. π and κ: paradoxes of kappa, *Optimal Data Analysis, 5*, 160-161.

Yıldıztekin, B. (2014). *Klasik test kuramı ve genellenebilirlik kuramından puanlayıcılar arası tutarlığın farklı yöntemlere göre karşılaştırılması.* Yüksek Lisans Tezi. Hacettepe Üniversitesi Eğitim Bilimleri Enstitüsü Eğitim Bilimleri Ana Bilim Dalı Eğitimde Ölçme ve Değerlendirme Bilim Dalı, Ankara.

Zapf, A., Castell, S., Morawietz, L. & Karch, A.(2016). Measuring inter-rater reliability for nominal data – which coefficients and confidence intervals are appropriate?. *Medical Research Methodology, 16* (93), 1-10.

Zhao, X., Feng, G. C., Liu J. S. & Deng K. (2018). We agreed to measure agreement redefining reliability dejustifes Krippendorff's alpha. *China Media Research, 14* (2), 1-15.