ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2018; 9(2);202-215

Otizm Sosyal Beceriler Profili Ölçeğinde Puanlayıcılar Arası Güvenirliğin Farklı Kuramlara Göre Karşılaştırılması*

Comparison of Interrater Reliability Based on Different Theories for Autism Social Skills Profile

Zeynep PEKİN ** Sevda ÇETİN *** Neşe GÜLER ****

Öz

Bu araştırmada, "Otizm Sosyal Beceriler Profili" (OSBP) ölçeğinin beş puanlayıcı tarafından puanlanması ile elde edilen puanların klasik test kuramı ve genellenebilirlik (G) kuramı ile puanlayıcılar arası güvenirliğinin karşılaştırılması amaçlanmıştır. G kuramında puanlayıcıların birlikte ve dönüşümlü puanlama yapmasıyla oluşturulan farklı desenlerden ve klasik test kuramından elde edilen güvenirlik katsayılarının düzeyleri saptanmış ve hangi kuramın daha fazla bilgi sunduğu belirlenmeye çalışılmıştır. Araştırmada elde edilen veriler klasik test kuramında her bir puanlayıcı için puanların iç tutarlılık güvenirliği Cronbach-alfa (α) katsayısı; puanlayıcılar arası güvenirlik, Kendall'ın uyuşum katsayısı, puanlayıcılar arası korelasyon katsayısı ve puanlayıcıların verdikleri puanlar arasında fark olup olmadığı ise ilişkili örneklemlerde varyans analizi ile hesaplanmıştır. Genellenebilirlik teorisinde, değerlendiricilerin ortaklaşa ve alternatif derecelendirmelerine göre iki farklı tasarım oluşturulmuştur. G kuramı kapsamında bireylerin (*b*) aynı maddeler (*m*) doğrultusunda puanlayıcıların (*p*) her biri tarafından puanlandığı *bxmxp* çapraz deseni ve bireylerin tüm maddeler doğrultusunda farklı puanlayıcılar tarafından puanlandığı (*p:b*)*xm* yuvalanmış deseni için ayrı ayrı G ve K çalışmaları yapılmış ve sonuçlar birbirleriyle karşılaştırılmıştır.

Anahtar Kelimeler: Klasik test kuramı, genellenebilirlik kuramı, puanlayıcılar arası güvenirlik, Kendall'ın uyuşum katsayısı, sosyal becerilerin değerlendirilmesi

Abstract

In this study, interrater reliability was compared based on both classical test theory and generalizability theory according to the scores which were obtained from five raters' ratings with Autism Social Skills Profile. Levels of reliability coefficients obtained from classical test theory and different designs in generalizability theory formed by five raters' jointly and alternatively ratings were determined and which theory presented more information was tried to be specified. In the classical test theory, Cronbach-Alpha (α) coefficient for internal consistency, Kendall's coefficient of concordance for inter-rater reliability and correlation coefficients of five raters' scores were calculated and it was investigated whether there was a difference among the means of raters' scores with F test. In the generalizability theory, two different designs were formed according to raters' jointly and alternatively ratings. Several G and D studies were made for crossed design pxixr (p: person, i: item and r: rater) which people were scored by all raters through all items and nested design (r:p)xi which people were scored by different raters through all items and the results were compared to each other.

Keywords: Classical test theory, generalizability theory, interrater reliability, Kendall's coefficient of concordance, evaluation of social skills

* Bu çalışma, Zeynep Pekin' in Hacettepe Üniversitesi, Eğitim Bilimleri Enstitüsünde, Dr. Öğr. Üyesi Sevda Çetin'in danışmanlığında hazırlanan yüksek lisans tezinden üretilmiştir.

** Arş. Gör., Yeditepe Üniversitesi, Eğitim Fakültesi, İstanbul-Türkiye, zynppknn@gmail.com:, http://orcid.org/0000-0002-9976-1218
*** Dr. Öğr. Üyesi, Hacettepe Üniversitesi, Eğitim Fakültesi, Ankara-Türkiye, tsevda@hacettepe.edu.tr http://orcid.org/0000-0001-5483-595X
**** Doç. Dr., İzmir Demokrasi Üniversitesi, Eğitim Fakültesi, İzmir-Türkiye, gnguler@gmail.com: https://orcid.org/0000-0002-2836-3132

Geliş Tarihi: 01.02.2018
DOI: *10.21031/epod.388590* *Kabul Tarihi: 18.05.2018*

GİRİŞ

Sosyal beceriler, sosyal bir ortamda kabul edilebilir bir şekilde başkalarıyla iletişime geçebilme yeteneğidir (Combs & Salaby, 1977). Bu becerilerde görülen yetersizlikler bireylerin sosyal ve eğitim hayatlarında iletişim kurmalarında sıkıntı yaşamalarına sebep olmaktadır. Bu durumdan yüksek düzeyde etkilenen özel gereksinim gruplarından birini de otizm spektrum bozukluğu olan bireyler oluşturmaktadır. Nitekim alanyazında otizm spektrum bozukluğu (OSB), sosyal etkileşimlerde yüksek seviyede yetersizlikler ve stereotip davranışlar ile karakterize nöro-gelişimsel bir bozukluk olarak tanımlanmaktadır (Özdemir, Diken, Diken & Şekercioğlu, 2013). Otizmli bireylerin sosyal becerilerindeki sınırlılıkları, başkalarına yaklaşmada sıra dışı özellikler gösterme (nerede duracağını bilememek), arkadaşlık kurmada sıkıntı yaşama, grup etkinliklerinde zorlanma, yalnızlığı yeğleme, başkalarının dikkatini çekme çabası göstermeme, sözel övgüler karşısında tepkisiz kalma, başkalarına karşı ilgisiz olma, başkalarının duygularını anlamada yetersiz olma şeklinde sıralanabilmektedir (Kırcaali-İftar, 2012).

Otizmli bireyler iletişim kurmada sıkıntı yaşamaları sebebiyle aile ve akranlarından kopuk bir yaşam sürmektedirler. Bu sebeple sosyal becerileri doğal ortamda gözlemleyerek arkadaş ve ailelerinden öğrenememektedirler. Bu becerilerin öğrenilebilmesi için bireyin özelliklerine göre yapılandırılmış sosyal beceri öğretim programlarına ihtiyaç duyulmaktadır (Hall & Schlesinger, 1997). Bu programların temelini ise otizmli bireylerin sosyal becerilerinin değerlendirilmesi oluşturmaktadır (Merrel, 2001). Ancak, özel eğitim alanında yapılan çalışmalarda çoğunlukla otizmli bireylerin kendilerini değerlendirecek yeterlilikte olmamaları sebebiyle değerlendirmeler aile, öğretmen gibi bireye yakın kişiler tarafından dereceleme ölçekleriyle yapılmaktadır. Bu tarz değerlendirmelerde, dikkatsizlik, kişisel yanlılık, halo etkisi, merkeze kayma etkisi, genelleme hatası, gözlem yetersizliği vb. gibi puanlayıcı kaynaklı hatalar karışabilmektedir (Turgut & Baykul, 2014). Bu sebeple, güFnilir sonuçların elde edilebilmesi için özel eğitim alanında değerlendirme yapılırken puanlayıcılar arası güvenirliğin test edilmesi oldukça önem kazanmaktadır. Puanlayıcılar arası güvenirliğin belirlenmesinde kullanılan çeşitli kuram ve uygulamaları bulunmaktadır. Ancak bu araştırmada, klasik test kuramı ve genellenebilirlik kuramı kapsamındaki yöntemler kullanıldığı için sadece bu kuramlara dair bilgiler yer almaktadır.

Klasik Test Kuramı

Klasik test kuramında (KTK), gerçek puan varyansı ve hata varyansının toplamından gözlenen puan varyansı oluşmaktadır (Lord & Novick, 1968). Gerçek varyans hariç kalan tüm varyansın farklı hata kaynaklarından gelebileceği düşünülmektedir. KTK'da, ele alınan hata kaynağına göre güvenirlik hesaplama yöntemleri farklılık göstermektedir (Baykul, 2000). Güvenirlik hesaplanırken olası hata kaynağı, test-tekrar test yönteminde ele alınan zaman, paralel (eşdeğer) formlar yönteminde formlar, iç tutarlılık anlamında hesaplama yapılırken ise görevler ya da maddeler olmaktadır (Shavelson & Webb, 1991; Brennan, 2001; Güler & Gelbal, 2010). Birden fazla puanlayıcının yer aldığı ölçmelerde ise puanlayıcılar hata kaynağı olarak düşünülmekte ve puanlayıcıların verdikleri puanlar arasındaki tutarlılığa 'puanlayıcılar arası güvenirlik' denilmektedir (Güler, 2008). Birden fazla puanlayıcının değerlendirme yaptığı çalışmalarda, puanlayıcılar önemli bir hata kaynağı olmakla birlikte çalışmadaki tek hata kaynağı değildir. Sonuçların kolay yorumlanması sebebiyle sıkça tercih edilmesine rağmen, KTK kapsamında yapılan güvenirlik yöntemlerinde hata kaynaklarının ayrı ayrı ele alınması önemli bir sınırlılık olarak karşımıza çıkmaktadır (Güler & Gelbal, 2010). Genellenebilirlik (G) kuramı, birden fazla hata kaynağının olduğu durumlarda güvenilirlik kestirimi için klasik test kuramının bir uzantısı olarak geliştirilen bir yaklaşımdır. G kuramı, tek bir analizle hata kaynaklarının aynı anda kestirimini sağlamakla birlikte, farklı uygulamalara katkı sağlayacak hata varyanslarına ilişkin bilgi edinilmesine de imkân tanımaktadır. (Cardinet, Johnson & Pini, 2009).

Genellenebilirlik Kuramı

Genellenebilirlik (G) kuramı, davranış değerlendirilmesinde güvenirliğin hesaplanmasını, güvenilir gözlemlerin genellenebilirlik (G) ve karar (K) çalışmalarıyla tasarlanıp, araştırılmasını ve çeşitli hata kaynaklarını göz önünde bulundurarak tek bir güvenirlik katsayısının bulunmasını sağlayan istatiksel bir kuramdır (Eason, 1989).

G kuramında; zaman, puanlayıcı, madde, formlar gibi benzerlik gösteren ölçme durumlarına, değişkenlik kaynaklarına yüzey (facet) adı verilir ve faktör analizinde yer alan faktörler gibi düşünülebilir (Güler, Uyanık & Teker, 2012). Her bir yüzeyin düzeyleri ise koşul olarak adlandırılmaktadır (Crocker & Algina, 1986; Shavelson & Webb, 1991; Brennan, 2001). Örneğin, "madde" bir yüzey olarak ele alınırsa; maddelerin her biri bir koşul olmaktadır. Araştırmacının, genellemek istediği yüzeyin koşullarına genelleme evreni, alınabilecek tüm koşulların oluşturacağı evrene ise kabul edilebilir gözlemler evreni denir (Crocker & Algina, 1986). Ölçme sonucunda, istenilen kararların alınacağı ölçmenin hedefi durumundaki değişkenlik kaynağı ise ölçme objesi olarak adlandırılmaktadır. G kuramında, ölçme objesine yüzey adı verilmemektedir. Ölçme objesi genellikle sistematik varyans içeren bireylerdir. Çünkü bireyler, doğası gereği farklılık gösterir. Ancak bireyler her zaman ölçmenin objesi olmak zorunda değildir. Ölçme durumuna göre, sistematik varyans içeren diğer madde, durum vb. de ölçme objesi olabilir (Eason, 1989; Shavelson & Webb, 1991; Mushquash & O'connor, 2006).

G kuramında, G çalışması ve K çalışması adı verilen iki tür çalışma yer almaktadır. G çalışmasının amacı, kabul edilebilir gözlemler evrenine ilişkin varyans bileşenlerini belirlemek ve bu varyans kaynakları hakkında bilgi edinmektir (Shavelson & Webb, 1991; Keiffer, 1998; Brennan, 2001). Varyans bileşenleri belirlendikten sonra 'Eğer ki?' sorularına cevap aramak için K çalışmasına geçilir. K çalışması sürecinde araştırmacı, madde, form ya da puanlayıcı sayıları gibi yüzeylerin koşulları üzerinde değişiklikler yaparak daha yüksek güvenirlik ve daha düşük hata içeren sonuçlar elde edebilecek senaryolar üretebilir (Kieffer, 1998).

Değişkenlik kaynaklarının ele alınış şekillerine göre G kuramında iki tür desen vardır. Bir ölçmedeki değişkenlik kaynaklarının (ölçme objesi ve yüzeyler dâhil) tüm koşulları, diğer değişkenlik kaynaklarının tüm koşullarıyla etkileşim gösteriyorsa bu desene çaprazlanmış desen denir ve bu desende değişkenlik kaynakları arasına *'x'* işareti konulur. Bir ölçmedeki bir değişkenlik kaynağının bazı koşulları, diğer bir değişkenlik kaynağının bazı koşullarıyla etkileşim gösteriyorsa; bu desene de yuvalanmış desen denir ve bu desende de değişkenlik kaynakları arasında *':'* işareti yer alır (Shavelson & Webb, 1991; Brennan, 2001; Mushquash & O'Connor, 2006). Ayrıca, G kuramı, mutlak ve bağıl olmak üzere iki tür değerlendirmeye olanak sağlar ve bu iki değerlendirme için iki farklı güvenirlik katsayısının hesaplanmasına imkân tanır. Bağıl değerlendirmeler için G-katsayısı, mutlak değerlendirmeler için ise Phi-katsayısı (Φ) hesaplanarak güvenirlik kestirilmektedir (Crocker & Algina, 1986; Shavelson & Webb, 1991; Brennan, 2001; Goodwin, 2001).

Alanyazında puanlayıcılar arası güvenirliğin KTK ve G kuramına göre karşılaştırıldığı çalışmalara rastlanılmaktadır (Rae & Hyland, 2001; Yelboğa & Tavşancıl, 2010; Öztürk, 2011; Deliceoğlu & Çıkrıkçı-Demirtaşlı 2012; Yıldıztekin, 2014, Polat-Demir, 2016). Örneğin, Rae ve Hyland (2001)'ın çalışmasında KTK ve G kuramı tutarlı sonuçlar vermiş ve puanlayıcılar arası güvenirlik yüksek bulunmuştur. Öte yandan, Öztürk (2011)'ün çalışmasında puanlayıcılar arası güvenirlik KTK ve G kuramı kapsamında düşük elde edilmiştir. Yelboğa ve Tavşancıl (2010)'ın çalışmalarında KTK'da puanlayıcılar arası güvenirlik için Kendall'ın uyuşum katsayısı ve G kuramında da G ve Phi katsayıları hesaplanmış ve her iki kuramdan elde edilen katsayıların birbirleriyle tutarlı oldukları sonucuna varılmıştır. Yıldıztekin (2014)'in çalışmasında KTK'da Pearson momentler çarpım korelasyon katsayısı, Spearman sıra farkları korelasyon katsayısı, Kappa ve Krippendorf Alfa katsayıları ile G kuramı karşılaştırılmış ve elde edilen sonuçlara göre puanlayıcılar arası güvenirlik yüksek bulunmuştur.

Genel olarak alanyazında puanlayıcılar arası güvenirliğin KTK ve G kuramına göre karşılaştırıldığı çalışmalar yer alsa da, özel eğitim alanında KTK ve G kuramının karşılaştırıldığı çalışmalara rastlanılmamıştır. Özel eğitim alanında otizmli bireylerin değerlendirilmesinin öğretmen gibi bireye

yakın kişiler tarafından yapıldığı göz önünde bulundurulduğunda, puanlayıcılar arası güvenirlik çalışmasının önemli olduğu düşünülmektedir. Bu alanda çalışan araştırmacıların, çalışmalarında puanlayıcılar arası güvenirliği genellikle KTK'ya dayalı yöntemlerle belirledikleri görülmektedir (Irmak, Sütçü, Aydın & Sorias, 2007; Girli & Atasoy, 2010; Sucuoğlu & Demir, 2017). Alanyazında birden fazla puanlayıcının değerlendirme yaptığı çalışmalarda G kuramı, KTK'ya göre daha kapsamlı ve güvenilir sonuçlar üretmesi nedeniyle daha çok önerilmektedir (Shavelson & Webb, 1991; Yelboğa & Tavşancıl, 2010; Güler & Gelbal, 2010; Güler, 2011). Bu doğrultuda gerçekleştirilen araştırmanın amacı, Otizm Sosyal Beceri Profili (OSBP) ile otizmli çocuk ve gençlerin birden fazla puanlayıcı tarafından puanlanması sonucu elde edilen sonuçlarda puanlayıcılar arası güvenirliğin KTK ve G kuramına göre karşılaştırılmasıdır. Bu araştırma ile kuramlardan elde edilen bilgilerin karşılaştırılmasına, gelecek araştırmalara ve özel eğitim alanında ölçme ve değerlendirme sürecine yönelik bir katkı sağlanabileceği düşünülmektedir. Ayrıca, bu çalışmada G kuramı kapsamında puanlayıcıların birlikte ve dönüşümlü puanlama yapmasıyla oluşturulan farklı desenlerin sonuçlarının karşılaştırılması amaçlanmaktadır. Her puanlayıcının her bireyi bütün maddeler doğrultusunda puanladığı çapraz desenler uygulama açısından zaman zaman çok pratik olamamaktadır. Bununla birlikte puanlayıcıların yorulması sebebiyle değerlendirmelerin güvenirliği riske girebilmektedir. Bu nedenle çalışmada her iki desenle elde edilen sonuçların karşılaştırılarak, yeterli düzeyde güvenilir sonuçlar için yuvalanmış desenin uygunluğunun belirlenmesinin gelecek çalışmalara katkı sağlayacağı düşünülmektedir.

Problem Cümlesi

Otizm Sosyal Beceriler Profili (OSBP) ölçeğinin sosyal karşılıklılık alt boyutunun birden fazla puanlayıcı tarafından puanlanması sonucu elde edilen puanlayıcılar arası güvenirliğin, KTK ve G kuramındaki farklı desen çalışmaları sonuçları nasıldır?

Alt Problemler

1. KTK'na göre; OSBP sosyal karşılıklılık alt boyutundan elde edilen puanların puanlayıcılara ilişkin iç tutarlılık düzeyi ve beş farklı puanlayıcı arasındaki tutarlılık derecesi nedir?
2. G kuramına göre; birey (*b*), madde (*m*) ve puanlayıcı (*p*) değişkenlerinin çaprazlandığı *bxmxp* deseninin sonuçları nasıldır?
3. G kuramına göre; birey (*b*) ve puanlayıcı (*p*) değişkenlerinin yuvalandığı, madde (*m*) değişkeninin ise çaprazlandığı *(p:b)xm* deseninin sonuçları nasıldır?
4. *bxmxp* ve *(p:b)xm* desenlerinden elde edilen G çalışması sonuçlarının karşılaştırılması nasıldır?

YÖNTEM

Araştırmanın Türü

Araştırma, puanlayıcılar arası güvenirliğin hesaplanmasında KTK ve G kuramından hangisinin daha çok bilgi sağladığını belirlemesi ve iki kuramdan elde edilen sonuçları karşılaştırması açısından karşılaştırmalı bir araştırmadır. Aynı zamanda araştırma, G kuramı ve KTK ile OSBP ölçeğine ait özelliklerin belirlenmesi yönüyle durum belirleme çalışması olduğundan betimsel bir araştırma niteliği taşımaktadır.

Araştırma Grubu

Araştırmanın çalışma grubunu, Ankara'da bir özel eğitim ve rehabilitasyon merkezinde eğitim almakta olan, "Otizm", "Asperger Sendromu", "Başka Şekilde Tanımlanamayan" tanılarını almış ya

da bu gruplardan herhangi birine dahil edilemeyen ancak otizm spektrum bozukluğu belirtileri gösteren "Yaygın Gelişimsel Bozukluk" tanısı almış 6-17 yaş arasında olan 50 çocuk ve genç oluşturmuştur. Örneklem grubu seçilirken, puanlayıcıların en az bir yıldır birlikte çalıştıkları öğrenciler seçilmiştir. 50 öğrenci, OSBP ölçeğinin "sosyal karşılıklılık" alt boyutunda bulunan 15 madde ile beş puanlayıcı tarafından puanlanmıştır. Puanlayıcı grubunu ise aynı kurumda 3-5 yıldır görev yapmakta olan iki özel eğitim öğretmeni, bir psikolog (özel eğitim alanıyla ilgilenen), bir fizyoterapist ve bir sosyal psikolog (aynı zamanda davranış terapisti) oluşturmaktadır.

Veri Toplama Aracı

OSBP, Bellini ve Hopf (2007) tarafından otizmli çocukların sosyal beceri yetersizliklerinin belirlenmesi ve bulgular doğrultusunda uygun müdahale programlarının oluşturulması gayesiyle geliştirilmiştir. Araştırmanın örneklemini, otizm spektrum bozukluğuna sahip 6-17 yaş arası 340 çocuk ve genç oluşturmuştur. 45 maddelik OSBP, sosyal karşılıklılık, sosyal katılım/kaçınma ve zarar verici sosyal davranışlar şeklinde adlandırılan üç alt boyuttan meydana gelmektedir. Bu çalışmada, OSBP ölçeğinin "sosyal karşılıklılık" alt ölçeğinde yer alan 15 madde kullanılmıştır. Ölçeğin Cronbach-alfa iç tutarlık katsayısı 0.92 iken, sosyal karşılıklılık alt ölçeği için yine bu değer 0.92 olarak bulunmuştur.

OSBP'nin Türkçe uyarlaması Demir (2009) tarafından 208 otizmli çocuk ve gencin ebeveynleri tarafından değerlendirilmesi sonucu elde edilen verilerle gerçekleştirilmiştir. Cronbach-alfa katsayıları alt ölçeklerden sosyal karşılıklılık için 0.91 ve ölçeğin tamamı için 0.84 bulunmuştur. Bu sonuçlara göre OSBP ölçeğinin geçerli ve güvenilir bir araç olduğuna karar verilmiştir (Demir, 2009).

Verilerin Analizi

Dörtlü likert tipi bir ölçek olan OSBP kullanılarak 50 öğrenci, beş uzman tarafından değerlendirilmiştir. Puanlayıcılar birbirlerinden bağımsız puanlama yapmışlardır. Elde edilen veriler, KTK ve G kuramındaki çaprazlanmış *bxmxp* (birey, madde, puanlayıcı) deseni ve birey ile puanlayıcıların yuvalandığı, maddelerin ise çaprazlandığı yuvalanmış desen olan *(p:b)xm* deseniyle analiz edilmiştir. Yapılan analizlerde SPSS 20.0 ve EduG 6.1 bilgisayar programlarından yararlanılmıştır.

BULGULAR

Çalışmada yer alan beş puanlayıcının ölçekte yer alan 15 madde doğrultusunda 50 bireye verdikleri puanların betimsel istatistikleri Tablo 1'de verilmiştir.

Tablo 1. Elde Edilen Puanların Beş Puanlayıcıya İlişkin Betimsel İstatistikleri

İstatistikler	***Puanlayıcılar***				
	1	***2***	***3***	***4***	***5***
Minimum	15	15	15	17	15
Maksimum	59	56	60	53	55
Ortalama	33.98	30.68	33.38	29.88	23.94
Std. Sapma	12.28	13.00	12.73	10.06	9.41
Çarpıklık	-0.023	0.312	-0.181	0.654	1.33
Basıklık	1.33	-1.25	-1.26	-0.437	1.44

Tablo 1. incelendiğinde, en yüksek ortalamanın birinci puanlayıcıya (33.98), en düşük ortalamanın ise beşinci puanlayıcıya (23.94) ait olduğu görülmektedir. Basıklık çarpıklık katsayılarının ise -1.5 ile +1.5 arasında değerler aldığı gözlemlenmektedir. Tabachnick ve Fidell (2013)'e göre bu katsayıların -1.5 ile +1.5 arasında olması verilerin normal dağılımına işaret etmektedir.

Birinci Alt Probleme İlişkin Bulgular

"KTK'ya göre; OSBP'nin sosyal karşılıklılık alt boyutundan elde edilen puanların puanlayıcılara ilişkin iç tutarlılık düzeyi ve beş farklı puanlayıcı arasındaki tutarlılık derecesi nedir?"

OSBP ölçeğindeki maddelerin kendi içinde tutarlı ölçme yapıp yapmadığının belirlenmesi amacıyla her bir puanlayıcı için iç tutarlılık güvenirliğini ifade eden Cronbach-alfa (α) katsayıları hesaplanmıştır. Ölçeğin "sosyal karşılıklılık" alt ölçeğinden elde edilen puanların her bir puanlayıcı için iç tutarlılığı Tablo 2'de verilmiştir.

Tablo 2. OSBP Ölçeği ile Yapılan Puanlamalara ait Cronbach-alfa (α) Değerleri

	Puanlayıcılar				
	1	***2***	***3***	***4***	***5***
Cronbach α	0.981	0.979	0.988	0.961	0.953

Tablo 2 incelendiğinde, her bir puanlayıcı için hesaplanan iç tutarlılık değerleri 0.95 ile 0.98 arasında değişmektedir. Daha sonra, puanlayıcılar arası tutarlılık derecesi, parametrik olmayan istatistiksel bir teknik olan Kendall'ın uyuşum katsayısı ile analiz edilmiştir. Analiz sonucunda uyuşum katsayısı 15 madde için 0.201 olarak bulunmuştur (X^2= 40.272, *sd*=4, *p*= .00 < . 05). Ayrıca, beş puanlayıcının 15 madde üzerinden verdikleri puanlar arasındaki korelasyon değerleri Tablo 3'te gösterilmiştir.

Tablo 3. Beş Puanlayıcının 15 Maddeye Verdikleri Puanlar Arasındaki Korelasyon Katsayıları

	1.Puanlayıcı	***2.Puanlayıcı***	***3.Puanlayıcı***	***4.Puanlayıcı***	***5.Puanlayıcı***
1.Puanlayıcı	-	0.606*	0.921*	0.722*	0.362*
2.Puanlayıcı		-	0.578*	0.727*	0.585*
3.Puanlayıcı			-	0.678*	0.398*
4.Puanlayıcı				-	0.535*
5.Puanlayıcı					-

**p<0.01*

Tablo 3 incelendiğinde, beş puanlayıcının 15 madde üzerinden verdiği puanlar arasındaki korelasyon katsayıları 0.362 ile 0.901 arasında değiştiği gözlenmektedir. Birinci ve üçüncü puanlayıcı, birinci ve dördüncü puanlayıcı son olarak da ikinci ve dördüncü puanlayıcı arasında yüksek; diğer puanlayıcılar arasında ise orta derecede anlamlı ilişki vardır.

Araştırmada korelasyon değerleri hesaplandıktan sonra, elde edilen puanların ortalamaları arasında farklılık olup olmadığı, ilişkili örneklemlerde tek faktörlü varyans analizi ile test edilmiş ve istatiksel olarak anlamlı bir farklılık bulunmuştur (*F*=8.261, p= .00<.05). Bu sonuç üzerine puanlayıcıların puan ortalamalarının ikili karşılaştırılması için çoklu karşılaştırma çalışması yapılmıştır. Çoklu karşılaştırma çalışmasında, beşinci puanlayıcı ile diğer tüm puanlayıcılar arasında ve birinci ile dördüncü puanlayıcı arasında anlamlı bir farklılık bulunmuştur.

İkinci Alt Probleme İlişkin Bulgular

"G kuramına göre; birey (b), madde (m) ve puanlayıcı (p) değişkenlerinin çapraz tasarlandığı bxmxp deseninin sonuçları nasıldır?"

İkinci alt problem için birey (*b*), madde (*m*) ve puanlayıcı (*p*) değişkenlerinin çapraz tasarlandığı *bxmxp* deseni ele alınmıştır. Bu aşamada ilk olarak, G çalışması sonucunda kestirilen varyans bileşenleri ve toplam varyansı açıklama yüzdeleri incelenmiştir. G çalışması sonucunda kestirilen

varyans bileşenleri ve toplam varyansı açıklama yüzdeleri *b, m* ve *p* ana etkileri ile *bm, bp, mp* ve *bmp* ortak etkileri Tablo 4'te verilmiştir.

Tablo 4. *bxmxp* Desenine ait G Çalışması Sonucunda Kestirilen Varyans Bileşenleri ve Toplam Varyansı Açıklama Yüzdeleri

Varyans Kaynağı	*sd*	*Toplam Kareler*	*Kareler Ortalaması*	*Varyans*	*%*
b	49	1529.890	31.222	0.36541	39.9
m	14	76.195	5.442	0.01675	1.8
p	4	210.878	52.719	0.06412	7.0
bm	686	217.937	0.317	0.02522	2.8
bp	196	723.254	3.690	0.23323	25.5
mp	56	63.289	1.130	0.01877	2.1
bmp	2744	525.777	0.191	0.19161	20.9
Toplam	3749	3347.223			100

b:birey, m: madde, p:puanlayıcı

Tabloda verilen G çalışması sonucunda kestirilen varyans bileşenleri ve toplam varyansı açıklama yüzdeleri incelendiğinde, en büyük oranda birey (*b*) ana etkisinin (%39.9), daha sonra birey *x* puanlayıcı (*bp*) ortak etkisinin (%25.5) toplam varyansı açıkladığı, en az ise %1.8 değer ile madde (*m*) ana etkisinin toplam varyansa katkı sağladığı görülmektedir.

Ana etkilere ait varyans bileşenleri ve toplam varyansı açıklama yüzdeleri incelendiğinde, birey (*b*) ana etkisi, toplam varyansın % 39.9'unu açıklamaktadır. Bu değer ile toplam varyansa en çok katkı sağlayan birey varyans bileşeni, bireylerin ölçülen özellik bakımından birbirlerinden farklılaştığını göstermektedir. Madde (*m*) ana etkisi ise toplam varyansın % 1.8'ini açıklamaktadır. Bu değer ile toplam varyansı açıklamada en son sırada yer almaktadır. Madde varyans bileşeninin toplam varyansı açıklama yüzdesinin düşük olması, her maddenin benzer güçlük düzeyinde olduğuna işaret etmektedir. Son olarak, puanlayıcı (*p*) ana etkisi, toplam varyansın % 7'sini açıklamakta olup; bu varyans bileşeni, ana etkilerden toplam varyansa en çok katkı sağlayan ikinci bileşen olarak gözlenmektedir. Bu durum, puanlayıcıların birbirleriyle çok benzer puanlama yapmadıklarını belirtmektedir.

Ortak etkilerden, birey *x* madde (*bm*) ortak etkisi, toplam varyansın % 2.8'ini açıklamaktadır. Birey *x* madde etkileşiminin toplam varyansa katkısının düşük olması, maddelerin güçlük düzeylerinin bireyden bireye farklılık göstermediğine işaret eder. Birey *x* puanlayıcı etkileşimi (*bp*) ölçülmek istenmeyen puanlayıcı etkisinin, ölçülmek istenen birey etkisini etkilemesi sonucu ortaya çıkan değişkenlik kaynağıdır. Birey x puanlayıcı ortak etkisi, toplam varyansın % 25.5'ini açıklayarak, toplam varyanstaki payı en yüksek ikinci değişkenlik kaynağıdır. Bu sonuç, bireylerin bağıl durumlarının bir puanlayıcıdan diğerine değiştiği anlamına gelmektedir. Madde x puanlayıcı (*mp*) ortak etkisi varyansı, % 2.1 toplam varyansı açıklama oranıyla, toplam varyans içindeki payı en düşük ikinci değişkenlik kaynağıdır. Toplam varyansı açıklama yüzdesinin düşük olması, maddelere verilen puanların puanlayıcıdan puanlayıcıya çok farklılaşmadığını, puanlayıcıların bireyleri maddelerin güçlük düzeyleri açısından tutarlı puanladıklarını göstermektedir.

Birey *x* madde *x* puanlayıcı (*bmp*) etkisine ait varyans bileşeni artık varyans olarak adlandırılıp, ölçme hatasını da barındırmaktadır. Tablo 4 incelendiğinde, ölçme objesi olan birey (*b*) ve birey x puanlayıcı (*bp*) ortak etkileşiminden sonra toplam varyansa en büyük getirisi olan üçüncü değişkenlik kaynağıdır. Artık varyans, toplam varyansın % 20.9'unu açıklamaktadır. Artık varyans bileşeninin yüksek çıkması; birey, madde ve puanlayıcı ortak etkisi ve/veya tesadüfi hataların büyük olabileceğinin bir göstergesidir.

OSBP ölçeğinin "sosyal karşılıklılık" alt boyutunda yer alan 15 madde için, beş puanlayıcının sayısının artırılıp azaltılarak puanlayıcı sayısının 3, 4, 6 ve 7 olduğu durumlara göre düzenlenen K çalışması senaryoları için kestirilen G ve Phi katsayılarına ilişkin değerler Tablo 5'te verilerek açıklanmıştır.

Tablo 5. *bxmxp* Desenine ait K Çalışmaları ile Puanlayıcı Sayılarının Artırılıp Azaltılmasıyla Yapılan Senaryolara Göre G ve Phi Katsayıları

Puanlayıcı Sayıları									
3		***4***		***5***		***6***		***7***	
G	***Φ***	***G***	***Φ***	***G***	***Φ***	***G***	***Φ***	***G***	***Φ***
0.813	0.774	0.852	0.819	**0.877**	**0.848**	0.895	0.869	0.908	0.885

Araştırmada kullanılan *bxmxp* deseninde, 50 bireyin beş puanlayıcı tarafından 15 madde doğrultusunda puanlanması ile elde edilen G katsayısı 0.877, Phi katsayısı ise 0.848 olarak bulunmuştur. Tablo 5'teki veriler incelendiğinde, puanlayıcı sayısı azaltıldığında G ve Phi katsayılarının azaldığı, puanlayıcı sayısı artırıldığında ise G ve Phi katsayılarının arttığı gözlemlenmektedir.

Üçüncü Alt Probleme İlişkin Bulgular

"G kuramına göre; birey (b) ve puanlayıcı (p) değişkenlerinin yuvalanmış, madde (m) değişkeninin ise çapraz tasarlandığı (p:b)xm deseninin sonuçları nasıldır?"

Üçüncü alt problem için birinci desende kullanılan aynı verilerle birey (*b*) ve puanlayıcı (*p*) değişkenlerinin yuvalanmış, madde (*m*) değişkeninin ise çaprazlandığı *(p:b)xm* deseni çalışılmıştır. Bu aşamada ilk olarak, G çalışması sonucunda kestirilen varyans bileşenleri ve toplam varyansı açıklama yüzdeleri incelenmiştir. Daha sonra K çalışmalarında, puanlayıcı sayılarının 3, 4, 6 ve 7 olduğu senaryolara ilişkin G ve Phi katsayılarının değişimine bakılmıştır. *b, m, p:b, bm* ve *mp:b* değişkenleri için G çalışması sonucunda kestirilen varyans bileşenleri ve toplam varyansı açıklama yüzdeleri Tablo 6'da verilmiştir.

Tablo 6. *(p:b)xm* Desenine ait G Çalışması Sonucunda Kestirilen Varyans Bileşenleri ve Toplam Varyansı Açıklama Yüzdeleri

Varyans Kaynağı	***Sd***	***Toplam Kareler***	***Kareler Ortalaması***	***Varyans***	***%***
b	4	1529.890	31.222	0.35259	39.1
m	14	76.195	5.442	0.02050	2.3
p:b	200	934.13	4.670	0.29735	33
bm	686	217.97	0.317	0.02146	2.4
mp:b	2800	589.06	0.210	0.21038	23.3
Toplam	3749	3347.223			100

b:birey, m: madde, p:puanlayıcı

Tablo 6'da verilen G çalışması sonucunda kestirilen varyans bileşenleri ve toplam varyansı açıklama yüzdelerine bakıldığında, en çok % 39.1 ile birey (*b*) ana etkisinin, daha sonra %33 ile birey ve puanlayıcı (*p:b*) ortak etkisinin toplam varyansı açıkladığı, en az ise %2.3 değer ile madde (*m*) ana etkisinin toplam varyansa katkı sağladığı görülmektedir. Bu sonuçlardan anlaşılacağı üzere, birey (*b*) ana etkisi, toplam varyansa en çok katkı sağlayan varyans bileşenidir. Bu durum, bireylerin ölçülen özellik bakımından farklılaştığına işaret etmektedir. Madde (*m*) ana etkisi, % 2.3 ile toplam varyansa getirisi en düşük olan bileşendir. Bu durum, ölçekte yer alan 15 maddenin zorluk-kolaylık düzeylerinin değişmediğini göstermektedir.

Her bir puanlayıcı, farklı bireyleri puanladığı için çalışmada birey değişkeniyle puanlayıcı değişkeni yuvalanmıştır. Buradaki $\sigma^2(b:p)$ varyans bileşeni, birey varyans bileşenini $\sigma^2(b)$ ve birey puanlayıcı ortak etkileşim varyans bileşenini $\sigma^2(bp)$ temsil etmektedir (Brennan, 2001). (*b:p*) için kestirilen

varyans değeri toplam varyansın % 33'ünü açıklayarak, toplam varyansa getirisi en yüksek ikinci değişkendir. Bu değerin yüksek olması, birey-puanlayıcı etkileşiminin farklılaştığı, bireylerin puanlarının bir puanlayıcıdan diğerine farklılık gösterdiğini belirtmektedir. Ortak etkilerden, birey x madde (*bm*) ortak etkisi ise toplam varyansın % 2.4'ünü açıklamaktadır. Birey x madde etkileşiminin toplam varyansa katkısının düşük olması, bireylerin bağıl durumlarının bir maddeden diğerine çok değişmediğini göstermektedir.

Tablo 6 incelendiğinde, ölçmenin objesi olan birey (*b*) ve birey puanlayıcı (*p:b*) etkileşiminden sonra toplam varyansa en büyük getirisi olan üçüncü değişkenlik kaynağı, artık varyans bileşenidir. Buradaki artık varyans bileşeni $\sigma^2(mp{:}b,e)$, madde puanlayıcı $\sigma^2(mp)$ varyans bileşenini ve birey madde puanlayıcı $\sigma^2(bmp,e)$ varyans bileşenini temsil etmektedir (Brennan, 2001). Artık varyans bileşeninin toplam varyansı açıklama yüzdesi % 23.3'tür. Artık varyans bileşeninin yüksek çıkması birey-madde-puanlayıcı ortak etkileşimi, madde-puanlayıcı ortak etkileşimi ve/veya tesadüfi hataların büyük olabileceğinin bir göstergesidir.

(p:b)xm yuvalanmış deseninde birey ölçme objesi olup; puanlayıcı sayılarının azaltılıp artırılmasıyla düzenlenen senaryolar için yapılan K çalışmaları sonucunda kestirilen G ve Phi katsayılarına ait değerler Tablo 7'de verilerek açıklanmıştır.

Tablo 7. *(p:b)xm* Desenine ait K Çalışmaları ile Puanlayıcı Sayılarının Artırılıp Azaltılmasıyla Yapılan Senaryolara göre G ve Phi Katsayıları

Puanlayıcı Sayıları									
3		***4***		***5***		***6***		***7***	
G	***Φ***	***G***	***Φ***	***G***	***Φ***	***G***	***Φ***	***G***	***Φ***
0.770	0.767	0.816	0.813	**0.846**	**0.844**	0.868	0.865	0.884	0.881

Farklı puanlayıcı senaryolarına göre G katsayısı, Phi katsayısına göre daha yüksek değerlere sahiptir. Düzenlenen senaryolarda puanlayıcı sayısı azaltıldığında G ve Phi katsayılarında azalma, puanlayıcı sayıları artırıldığında ise G ve Phi sayılarında artış görülmektedir.

Dördüncü Alt Probleme İlişkin Bulgular

"bxmxp ve (p:b)xm desenlerinden elde edilen G çalışması sonuçlarının karşılaştırılması nasıldır?

Birey (*b*), madde (*m*) ve puanlayıcı (*p*) değişkenlerinin çaprazlandığı *bxmxp* deseni ile birey (*b*) ve puanlayıcı (*p*) değişkenlerinin yuvalanmış, madde (*m*) değişkeninin ise çaprazlanmış olduğu *(p:b)xm* deseninden elde edilen G çalışması sonuçları karşılaştırıldığında, bireylere ait varyans bileşeninin çapraz desende toplam varyansın % 39.9'unu, yuvalanmış desende ise toplam varyansın % 39.1'ini açıkladığı görülmektedir. Her iki desende de toplam varyansa en çok getirisi olan birey varyans bileşeni, bireylerin sosyal becerileri bakımından farklılaştığını göstermektedir.

Çapraz desende, puanlayıcı (*p*) varyans bileşeninin toplam varyansı açıklama oranı % 7, birey x puanlayıcı (*bp*) ortak etkileşimine ait varyans bileşeninin ise toplam varyansı açıklama oranı % 25.5'tir. Yuvalanmış desende ise puanlayıcı ana etkisine ve birey x puanlayıcı ortak etkisine ait varyans bileşenleri ayrı ayrı değil, (*b:p*) değişkeni altında ortak değerlendirilmektedir. (*b:p*) değişkeninin toplam varyansı açıklama oranı ise % 33'tür.

bxmxp ve *(p:b)xm* desenlerinde puanlayıcı sayılarının artırılıp azaltılmasıyla yapılan K çalışmaları karşılaştırıldığında, her iki desende de puanlayıcı sayıları artırıldığında G ve Phi katsayıları artarken, puanlayıcı sayıları azaltıldığında G ve Phi katsayıları azalmaktadır. Ancak, çapraz desende kestirilen G ve Phi katsayılarının yuvalanmış desende kestirilen G ve Phi katsayılarından daha yüksek olduğu görülmektedir.

SONUÇLAR ve TARTIŞMA

KTK kapsamında ilk olarak iç tutarlılık düzeyinin belirlenmesi için Cronbach-alfa güvenirlik katsayısı oldukça yüksek (α=0.95 - α=0.98) çıkmıştır. Cronbach-alfa değerlerinin 0.90'dan yüksek olması ölçeğin iç tutarlılığın oldukça yüksek olduğunu göstermektedir. Puanlamada kullanılan maddelerin OSBP ölçeğinin sosyal karşılıklılık alt boyutunda yer alan maddelerin tek bir yapıyı ölçmesi, bu değerlerin oldukça yüksek çıkmasını açıklar niteliktedir. Elde edilen Cronbach-alfa katsayılarına göre, puanlayıcıların verdiği puanların kendi içinde tutarlı olduğu yorumu yapılabilir. Beş puanlayıcı arasındaki uyum düzeyinin belirlenmesi için hesaplanan Kendall'ın uyuşum katsayısı, beklenen düzeyden (en az 0.80) düşük bulunmuştur. Bu bulgu, dereceleme ölçekleriyle gerçekleştirilen Deliceoğlu ve Çıkrıkçı-Demirtaşlı (2012)'nın ve Öztürk (2011)'ün çalışmalarında elde ettikleri düşük düzeydeki Kendall'ın uyuşum katsayıları bulgularıyla tutarlılık göstermektedir. Howell (2009)'e göre elde edilen bu sonuç doğrultusunda, puanlayıcılar arası uyum olmadığı yorumu yapılabilir. Ayrıca, her bir puanlayıcının verdiği puanlar ile diğer puanlayıcıların verdikleri puanlar arasındaki ilişkiler, Pearson momentler çarpım korelasyon katsayısı ile hesaplanmıştır. Ancak, Pearson çarpım moment korelasyon katsayısının ortalamadan bağımsız olması sebebiyle bu katsayı, puanlayıcıların verdikleri puanlar arasındaki benzerlik ve farklılıklar hakkında bilgi verememektedir. Bu nedenle, Goodwin (2001) puanlayıcılar arası tutarlılık test edilirken, korelasyonla birlikte ortalamaların da karşılaştırılmasını önermektedir. Bu doğrultuda gerçekleştirilen tek faktörlü varyans analizi sonucu elde edilen anlamlı farklılık, Kendall'ın uyuşum katsayısı ile elde edilen puanlayıcıların puanlamalarının birbirleriyle paralellik göstermediği yorumunu desteklemektedir.

G kuramı kapsamında iki farklı desen ile çalışılmıştır. Bunlardan ilki birey, madde ve puanlayıcı değişkenlerinin çaprazlandığı *bxmxp* desenidir. G çalışması sonucunda kestirilen varyans bileşenleri ve toplam varyansı açıklama yüzdeleri incelendiğinde, toplam varyansı açıklama yüzdesi en yüksek olan değişkenlik kaynağın birey (*b*) ana etkisi olduğu görülmüştür. Birey ana etkisinin, ölçmenin objesi olması nedeniyle bu durum istenilen bir durumdur (Güler vd., 2012). Birey puanlayıcı (*bp*) ortak etkisine ait varyans bileşeni, toplam varyansa katkısı en yüksek olan ikinci değişkenlik kaynağıdır. Bu değerin yüksek olması, bireylerin bağıl durumlarının bir puanlayıcıdan diğerine değiştiğini göstermiştir. Bu nedenle puanlayıcılar arası uyumun düşük olduğu yorumu yapılabilir. Puanlayıcı (*p*) değişkenine ait varyans bileşeninin, ana etkiler arasında ölçme objesinden sonra en yüksek varyansa sahip bileşen olması, bu durumu destekler niteliktedir. G kuramı kapsamında çalışılan bir diğer desen, birey ve puanlayıcıların birbiriyle yuvalandığı, maddelerin ise her ikisiyle çaprazlandığı *(p:b)xm* yuvalanmış desenidir. Bu desende yapılan G çalışması sonucunda toplam varyansa katkı sağlayan beş varyans bileşeninden, yüzdesi en yüksek olan birey (*b*) değişkenine ait varyans bileşenidir. Birey değişkeniyle puanlayıcı değişkeninin yuvalandığı (*p:b*) değişkenine ait varyans değeri, toplam varyansa en çok katkıyı sağlayan ikinci bileşendir. Bu değerin yüksek olması, birey-puanlayıcı etkileşiminin farklılaştığı, bireylerin puanlarının bir puanlayıcıdan diğerine farklılık gösterdiğini belirtmektedir. *bxmxp* çapraz ve *(p:b)xm* yuvalanmış desenlerinde yapılan G çalışmalarıyla elde edilen varyans ve toplam varyansı açıklama yüzdeleri karşılaştırıldığında genel olarak, çapraz ve yuvalanmış desenden elde edilen bulgulara göre kestirilen varyans değerleri arasında paralellik olduğu görülmüştür. Bu bulgu, Alkan (2013)'ün ve Nalbantoğlu (2009)'un çalışma bulguları ile desteklenmektedir. Her iki desende elde edilen sonuçlar birbirleriyle tutarlılık gösterse de, çapraz desenin, birey-puanlayıcı ve madde-puanlayıcı ortak etkileri için de bilgi üretmesi sebebiyle daha detaylı bilgi verdiği yorumu yapılabilir.

KTK ve G kuramı kapsamında yapılan analizler sonucu puanlayıcılar arası uyumun düşük çıkmasının nedenlerinin araştırılması için puanlayıcılarla görüşmeler yapılmıştır. Görüşmelerden elde edilen bilgiler doğrultusunda; puanlayıcıların mesleki alanlarının, sosyal becerileri farklı şekilde puanlamalarına neden olduğu sonucuna varılmıştır. Örneğin, aralarında yüksek ilişki bulunan birinci puanlayıcı (psikolog) ve üçüncü puanlayıcı (sosyal psikolog, davranış terapisti) değerlendirmelere psikolog olarak yaklaştıklarını ve bunun puanlamalara yansımış olabileceğini ifade etmişlerdir. Ayrıca, dereceleme ölçeklerinde güvenirliği tehlikeye sokan dikkatsizlik, kişisel yanlılık, merkeze kayma etkisi, halo etkisi gibi unsurların puanlayıcılar arası uyumu etkilemiş olabileceği

düşünülmektedir. Bu sebeple, farklı puanlayıcıların yer aldığı çalışmalar düzenlenirken, puanlayıcıların nasıl puanlama yapacaklarına dair bilgilendirme eğitimleri verilerek, ön bir uygulama ile puanlayıcıların farklı anladıkları ölçütler/maddeler ve nedenleri araştırılıp, bu farklılıklar giderilerek asıl uygulama gerçekleştirilebilir.

KTK'da sadece bağıl değerlendirmeler için güvenirlik hesaplarken; G kuramında hem bağıl değerlendirme için G-katsayısı hem de mutlak değerlendirmeler için Phi-katsayısı kestirilmektedir (Crocker & Algina, 1986; Shavelson & Webb, 1991; Brennan, 2001; Goodwin, 2001). Düzenlenen farklı puanlayıcı senaryolarında çapraz desende kestirilen G ve Phi katsayıları, yuvalanmış desende kestirilen G ve Phi katsayılarından daha yüksek kestirilmiştir. Ayrıca, gerek çapraz desende gerekse yuvalanmış desende tüm senaryo durumlarına göre G katsayıları Phi katsayılarından daha yüksek çıkıştır. Benzer çalışmalarda (Nalbantoğlu, 2009; Alkan, 2013) da G katsayılarının Phi katsayılarından yüksek kestirildiği görülmektedir. Bu durum, mutlak ve bağıl hata varyansları arasındaki farklılıktan kaynaklandığı için beklenen bir durumdur. Tanımı gereği mutlak hata varyansı, bağıl hata varyansından daha yüksek bir değere sahip olduğu için, Phi katsayısı G katsayısına göre daha düşük bir değere sahip olur. Shavelson ve Webb (1991)'e göre "yüksek" güvenirlik sağlayabilmek için G ve Phi katsayılarının en az 0.80 olması gerektiğini belirtmişlerdir. Yapılan K çalışmalarına göre, bu değerin sağlanabilmesi için her iki desende de puanlayıcı sayısının en az dört olması gerekmektedir. Bu durumda, zaman ve işgücü göz önünde bulundurularak yüksek güvenirlikli en ekonomik uygulamanın dört puanlayıcı ile yapılabileceği söylenebilir.

Elde edilen sonuçlar incelendiğinde, klasik test kuramında puanlayıcılar arası güvenirliğin belirlenebilmesi için birden fazla analize gerek duyulmaktadır. Genellenebilirlik kuramı ise birden fazla hata kaynağını aynı anda göz önünde bulundurarak güvenirlik çalışmalarının tek bir analiz ile yapılmasını sağlamaktadır. Ayrıca, klasik test kuramı kapsamında hesaplanan Kendall'ın uyuşum katsayısı toplam puanlar üzerinden analiz yaparken, G kuramı maddeler bazında analiz yaptığı için daha detaylı bilgi vermektedir. Bu sebeple, puanlayıcılar arası güvenirlik analizlerinde G kuramı tercih edilebilir. KTK'da sadece bağıl değerlendirmeler için güvenirlik hesaplanırken; G kuramında hem bağıl değerlendirme hem de mutlak değerlendirmeler için güvenirlik katsayıları kestirilmektedir. Bununla birlikte, G kuramı karar çalışmaları ile farklı senaryolar için güvenirlik kestirimleri yaparak, daha sonraki çalışmalar için daha yüksek güvenirlik ve daha düşük hata içeren sonuçlar elde edebilecek senaryolar üretebilmektedir. Bu sebeplerden ötürü, G kuramının KTK'na göre daha avantajlı bir kuram olduğu yorumu yapılabilir.

KAYNAKÇA

Alkan, M. (2013). *PISA 2009 okuma becerileri açık uçlu sorularının puanlanmasında genellenebilirlik kuramındaki farklı desenlerin karşılaştırılması* (Doktora Tezi, Hacettepe Üniversitesi, Ankara). https://tez.yok.gov.tr/UlusalTezMerkezi/ adresinden edinilmiştir.

Baykul, Y. (2000). *Eğitimde ve psikolojide ölçme: Klasik test teorisi ve uygulaması*. Ankara: ÖSYM.

Bellini, S., & Hopf, A. (2007). The development of the autism social skills profile: A preliminary analysis of psyhometric properties. *Focus on Autism and Other Developmental Disabilities, 22*(2), 80–87.

Brennan, R. L. (2001*). Generalizability theory*. New York: Springer-Verlog.

Cardinet, J., Johnson, S., & Pini, G. (2009). *Applying generalizability theory using eduG (quantitative methodology series)*. New York, London: Routledge.

Combs, H. L., & Slaby, D. A. (1977). Social skills training with children. In B. B. Lahey & A. E. Kazdin (Eds.), *Advances in clinical childpsy- chology* (Vol. 1, pp. 161-201). New York: Plenum.

Crocker, L. M., & Algina, L. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart and Winston.

Deliceoğlu, G. ve Çıkrıkçı-Demirtaşlı, N. (2012). Futbol yetilerine ilişkin dereceleme ölçeğinin güvenirliğinin genellenebilirlik kuramına ve klasik test kuramına dayalı olarak karşılaştırılması. *Spor Bilimleri Dergisi, 23*(1), 1-12.

Demir, Ş. (2009). *Otizmli çocukların sosyal becerilerinin farklı değişkenler açısından değerlendirilmesi* (Yüksek Lisans Tezi, Ankara Üniversitesi, Ankara). https://tez.yok.gov.tr/UlusalTezMerkezi/ adresinden edinilmiştir.

Eason, S. H. (1989, November). *Why generalizability theory yields beter results than classical test theory.* Mid-South Educational Research Association Annual Meeting. Little Rock, AR, USA

Goodwin, L. D. (2001). Interrater agreement and reliability. *Measurement in Psychical Education and Exercises Science, 5*(1), 13-34.

Girli, A., & Atasoy, S. (2010). Examining the effectiveness of social skills training program based on cognitive process approach among inclusion students with autism. *Elementary Education Online, 9*(3), 990-1006.

Güler, N. (2008). *Klasik test kuramı genellenebilirlik kuramı ve Rasch modeli üzerine bir araştırma* (Doktora Tezi, Hacettepe Üniversitesi, Ankara). https://tez.yok.gov.tr/UlusalTezMerkezi/ adresinden edinilmiştir.

Güler, N. ve Gelbal S. (2010). Açık uçlu matematik sorularının güvenirliğinin klasik test kuramı ve genellenebilirlik kuramına göre incelenmesi. *Kuram ve Uygulamada Eğitim Bilimleri Dergisi, 10*(2), 989–1019.

Güler, N. (2011). Rasgele veriler üzerinde genellenebilirlik kuramı ve klasik test kuramına göre güvenirliğin karşılaştırılması. *Eğitim ve Bilim, 36*(162), 225-234.

Güler, N., Uyanık, G. K. ve Teker, G. T. (2012). *Genellenebilirlik kuramı*. Ankara: Pegem Akademi.

Hall, J. A., Schlesinger, D. J., & Dieen, J. P. (1997). Social skills training in group with developmentally disabled adults. *Research on Social Work Practice, 7*(2), 187-201.

Howell, D. C. (2009). *Statistical methods for psychology.* (7th Edition). USA: Thomson Learning Academic Research Center.

Irmak, T. Y., Sütçü, S. T., Aydın, A. ve Sorias, O. (2007). Otizm davranış kontrol listesinin (abc) geçerlik ve güvenirliğinin incelenmesi. *Çocuk ve Gençlik Ruh Sağlığı Dergisi, 14*(1), 13-23

Kieffer, K. M., (1998,April). *Why generalizability theory is essential and classical test theory is often inadequate?* Paper presented at the annual meeting of the Southwestern Psychological Association, New Orleans, LA, USA.

Kırcaali-İftar, G. (2012). Otizm spektrum bozukluğuna genel bakış. E. Tekin-İftar (Ed). *Otizm spektrum bozukluğu olan çocuklar ve eğitimleri* (s. 17-46). Ankara: Vize.

Lord, F. M., & Novick, R. M. (1968). *Statistical theories of mental test scores.* California: Addison-Wesley Publishing Company.

Mushquash, C., & O'Connor, B. P. (2006). SPSS and SAS programs for generalizability theory analyses. *Behavior Research Methods, 38*(3), 542-547.

Nalbantoğlu, F. (2009). *Performans ölçümlerinde genellenebilirlik kuramıyla farklı desenlerin karşılaştırılması* (Yüksek Lisans Tezi, Hacettepe Üniversitesi, Ankara). https://tez.yok.gov.tr/UlusalTezMerkezi/ adresinden edinilmiştir.

Özdemir, O., Diken, İ. H., Diken, Ö. ve Şekercioğlu, G. (2013). Otizm davranış kontrol listesi (autism behavior checklist-ABC) modifiye edilmiş Türkçe versiyonunun geçerlik ve güvenirlik çalışması: Pilot uygulama sonuçları. *International Journal of Early Childhood Special Education (INT-JECSE), 5*(2), 168-186.

Öztürk, M. E. (2011). *Voleybol becerileri gözlem formu ile elde edilen puanların genellenebilirlik ve klasik test kuramına göre karşılaştırılması* (Doktora Tezi, Hacettepe Üniversitesi, Ankara). https://tez.yok.gov.tr/UlusalTezMerkezi/ adresinden edinilmiştir.

Polat-Demir, B. (2016). Vee diyagramından elde edilen puanların güvenirliğinin klasik test kuramı ve genellenebilirlik kuramına göre incelenmesi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 7*(2), 419-431.

Rae, G., & Hyland, P. (2001). Generalisability and classical test theory analyses of Koppitz's Scoring System for human figure drawing. *British Journal of Educational Psychology, 71*, 369-382.

Shavelson, R. J., & Webb, N. M. (1991*). Generalizability theory: A primer*. New-York: Springer.

Sucuoğlu, B. ve Demir, Ş. (2017). Bağlamsal Değerlendirme Envanteri: Otizmli bireylerin problem davranışlarının bağlamsal değişkenleri. *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Özel Eğitim Dergisi, 8*(2), 209-224.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Boston: Pearson.

Turgut, M. F., & Baykul, Y. (2014). *Eğitimde ölçme ve değerlendirme* (6. Baskı). Ankara: Pegem Akademi.

Yelboğa, A. ve Tavşancıl, E. (2010). Klasik test ve genellenebilirlik kuramına göre güvenirliğin bir iş performansı ölçeği üzerinde incelenmesi. *Kuram ve Uygulamada Eğitim Bilimleri, 10*(3), 1825-1854.

Yıldıztekin, B. (2014). *Klasik test teoremi ve genellenebilirlik kuramından puanlayıcılar arası tutarlılığın farklı yöntemlere göre karşılaştırılması* (Yüksek Lisans Tezi, Hacettepe Üniversitesi, Ankara). https://tez.yok.gov.tr/UlusalTezMerkezi/ adresinden edinilmiştir.

EXTENDED ABSTRACT

Introduction

The term social skills refers to the ability to communicate with others in an acceptable way in a social environment (Combs & Salaby, 1977). Inadequacies in these skills cause people to experience challenges in communicating in social and educational life. One of the special groups with special needs that are highly affected by this situation is those with autism spectrum disorder (Özdemir, Diken, Diken & Şekercioğlu, 2013). In the studies conducted in special education, the evaluations of the autistic people are usually done by scales of their families and teachers, because they are not competent enough to evaluate themselves. In such evaluations, rater-based errors can be encountered such as carelessness, personal bias, etc. (Turgut and Baykul, 2014). For this reason, it is very important to test the reliability between the raters while doing evaluation in the field of special education so that reliable results can be obtained.

Method

In this study, interrater reliability was compared based on classical test theory (CTT) and generalizability theory (GT) according to the scores which were obtained from five raters' ratings with Autism Social Skills Profile (ASSP). Levels of reliability coefficients obtained from CTT and different designs in GT formed by five raters' jointly and alternatively ratings were determined and which theory presented more information was tried to be specified. The research group consisted of 50 children and youths with autism who were being trained in a special education and rehabilitation center in Ankara and five raters rated them through social reciprocity sub-scale under Autism Social Skills Profile. The raters scored independently. The obtained data were analyzed in GT with the crossed design *pxixr* (person, item, rater) in which people were scored by all raters through all items and with the nested design *(r:pxi)* in which people were scored by different raters through all items. Additionally, Cronbach Alpha (α) coefficient for internal consistency, Kendall's concordance coefficient for interrater reliability and correlation coefficients of five raters' scores were calculated in CTT and it was investigated whether there was a difference among the means of raters' scores with F test. SPSS 20.0 and EduG 6.1 computer programmes were used in the analyses.

Results and Discussion

According to the obtained Cronbach Alpha coefficients (α = 0.95 - α = 0.98), it can be interpreted that the scores given by the raters are internally consistent. The Kendall's coefficient of concordance was found to be lower than the expected level (at least 0.80). This finding is consistent with the low level Kendall's coefficient of concordance findings of the studies conducted by Deliceoğlu and Çıkrıkçı-Demirtaşlı (2012) and Öztürk (2011) with grading scales. In addition, Pearson moments product correlation coefficients were found between 0.362 and 0.901. The significant difference obtained from the one-factor variance analysis supports the interpretation that the scores of the raters obtained by Kendall's coefficient of concordance are not parallel to each other.

Two different designs were used for the GT. In the *pxixr* design; it was found that the variance source having the highest total variance explanatory percentage is the main effect person (*p*). The variance component of the common effect; person x rater (*pr*) is the second variance source that brings the greatest contribution to the total variance. The high level of this value indicates that the relative status of the persons varies by raters. For this reason, it can be interpreted that the concordance between the raters is low. In the nested *(r:p)xi* design, the variance component of the variable person (*p*) has the highest percentage contributing to the total variance among five variance components. The variance value of the variable in which the variables of the rater and person (*r:p*) are nested is the second component that brings the greatest contribution to the total variance. The high level of this value indicates that the person-rater interaction differentiates, and that scores differ from one rater to another. Comparing the variance and total variance explanatory percentages obtained from the G studies conducted in both designs, it was generally found that the variance

values estimated according to the findings obtained from the crossed and nested designs are parallel with each other. This finding supports the findings of Alkan (2013) and Nalbantoğlu (2009).

G and Phi coefficients estimated in the crossed design in different rater scenarios are estimated much higher than those estimated in the nested design. In addition, G coefficients are found higher than the Phi coefficients in all scenarios regardless of the crossed or nested design. In similar studies (Nalbantoğlu, 2009; Alkan, 2013), it can be observed that the G coefficients are estimated higher than the Phi coefficients. Since the absolute error variance by definition has a value higher than the relative error variance, Phi coefficient has a lower value than G coefficient. According to Shavelson and Webb (1991), G and Phi coefficients should be at least 0.80 in order to provide "high" reliability. According to D studies, the number of raters on both designs must be at least four in order to achieve this value. In this case, it can be said that the most economical application with high reliability considering the time and labour force can be done with four raters.

When the obtained results are examined, in CTT multiple analyses are required in order to determine the inter-rater reliability. GT, on the other hand, ensures reliability studies with a single analysis taking into account multiple error sources at the same time. In addition, while the reliability is calculated only for the relative evaluations in the CTT; reliability coefficients for both relative and absolute evaluations are estimated in GT. Moreover, GT makes reliability estimations for different scenarios with decision-making studies and in this way, it can produce scenarios including higher reliability and lower errors for further studies. For these reasons, it can be said that GT is more advantageous theory than CTT.