# EXAMINATION OF SCORING RELIABILITY ACCORDING TO GENERALIZABILITY THEORY IN CHECKLIST, ANALYTIC RUBRIC AND RATING SCALES [1]

***Mehtap AKTAŞ***
*Arş. Gör., Mersin Üniversitesi, mhtpaktas@gmail.com*
*ORCID: 0000-0002-3192-7445*

***Devrim ALICI***
*Doç. Dr., Mersin Üniversitesi, devrimo.alici@gmail.com*
*ORCID: 0000-0001-5542-0609*

*Received: 22.06.2017* *Accepted: 29.08.2017*

**ABSTRACT**

The aim of this research is to examine the inter-rater reliability in the context of G theory when the same performance tasks are rated by different raters with the help of a checklist, rating scale and analytical rubric. To this end, a checklist, rating scale and analytic rubric were prepared to rate the story-writing skills of fifth grade students. Six stories selected from the stories written by the 5th grade students of the primary school were rated 45 different raters with three different scoring keys at intervals of 10-15 days. 100 samples each were drawn with 2, 3, 5 and 10 raters from 45 raters participating in the study. For the 400 samples obtained, reliability between the raters was calculated according to G theory. For the 100 samples obtained for each case, the median and standard error were calculated. When the median values of the reliability estimates are examined, the median values increase as the number of raters and the number of categories increase, except for the median of the reliability of the raters that the 5 raters make using the checklist; it was observed that the standard errors obtained decreased as the number of raters increased. It has been determined that the lowest standard error values are obtained in the case of 10 raters. When the number of raters was 5 and the number of category was 2, it was determined that the reliability estimation gave the highest value.

**Keywords:** Generalizability theory (GT), inter-rater reliability, checklist, rating scale, analytic rubric.

[1] Bu çalışma, Mehtap AKTAŞ tarafından Mersin Üniversitesi, Eğitim Bilimleri Enstitüsü'nde Doç. Dr. Devrim ALICI danışmanlığında yapılan yüksek lisans tez çalışmasının bir bölümüdür.

# KONTROL LİSTESİ, ANALİTİK RUBRİK VE DERECELEME ÖLÇEKLERİNDE PUANLAYICI GÜVENİRLİĞİNİN GENELLENEBİLİRLİK KURAMINA GÖRE İNCELENMESİ

**ÖZ**

Bu araştırmanın amacı, aynı performans görevlerinin farklı sayıda puanlayıcı tarafından kontrol listesi, dereceleme ölçeği ve analitik rubrik yardımıyla puanlanması durumunda, puanlayıcılar arası güvenirliklerinin G kuramı çerçevesinde incelenmesidir. Bu amaç doğrultusunda, 5. sınıf düzeyindeki öğrencilerin hikâye yazma becerilerini puanlamak amacıyla, kontrol listesi, dereceleme ölçeği ve analitik rubrik hazırlanmıştır. İlköğretim 5. sınıf öğrencilerine yazdırılan hikâyeler arasından seçilen 6 hikâye 45 puanlayıcıya üç farklı puanlama anahtarı ile 10-15 gün aralıklarla puanlattırılmıştır. Araştırmaya katılan 45 puanlayıcı içerisinden 2, 3, 5 ve 10 puanlayıcılı 100'er örneklem çekilmiştir. Elde edilen 400 örneklem için G kuramı'na göre puanlayıcılar arası güvenirlikleri hesaplanmıştır. Elde edilen 1200 hesaplamanın her bir durum için elde edilen 100 örneklemi için ortancaları ve standart hataları hesaplanmıştır. Güvenirlik kestirimlerinin ortanca değerleri incelendiğinde, 5 puanlayıcının kontrol listesi kullanarak yaptıkları puanlamaların güvenirliklerinin ortanca değeri hariç olmak üzere, puanlayıcı sayısı ve aynı zamanda kullanılan ölçeğin kategori sayısı arttıkça ortanca değerlerinin de arttığı; elde edilen standart hataların, puanlayıcı sayısı arttıkça azaldığı gözlenmiştir. En düşük standart hata değerlerinin, 10 puanlayıcı olması durumunda elde edildiği saptanmıştır. Puanlayıcı sayısı 5 ve kategori sayısı 2 olduğunda, güvenirlik kestiriminin en yüksek değeri verdiği belirlenmiştir.

**Anahtar Kelimeler:** Genellenebilirlik kuramı, puanlayıcılar arası güvenirlik, kontrol listesi, dereceleme ölçeği, analitik rubrik.

**EXTENDED SUMMARY**

**Introduction**

A significant part of the measurement and evaluation activities in education consists of studies to be able to make decisions about the students. Measurement evaluation studies are used to determine whether or not students have the knowledge and skills necessary to succeed in a class, their learning deficiencies, inaccuracies, learning levels, and guidance activities for the students in various topics.

Different measurement and evaluation tools and methods are used to gather information about students and make various decisions. These tools can be listed as measuring instruments consisting of long answer, short answer, true false and multiple choice items, activities such as performance tasks, projects, development files and checklists, rating scales, grade scoring keys.

Each measuring tool may be more appropriate in different situations than in others at different levels of acquisition. It is expected that the educators who will prepare a measurement tool, will first take into consideration their objective and the measured characteristic, and then student characteristics, duration etc. criteria. For example, checklists, rating scales and rubrics are usually used in performance measurements.

Intra-rater reliability and inter-rater reliability examinations may also be needed when reliability works are being conducted regarding measurements of performance duties. Rater reliability studies can be conducted via Generalizability theory (GT).

**Method**

The study was conducted on 104 students in grade 5 and 104 teachers in various schools. The students comprise the group that is completing the performance task assigned to them, and the teachers comprise the group of raters evaluating the performance task. Because of the participant losses experienced in the rater group in the study period, the research was conducted from the data obtained from 45 raters (teachers).

In the course of preparing and implementing the performance task, a checklist was prepared and the writing ability level of 5th grade Turkish lessons was listed and 7 class teachers were asked to decide which of these achievements could be observed with the help of the performance task. Using the simple fit index, achievements with 70% (Erkuş, 2006) or more fitness were determined. Examples of performance tasks were requested from the teachers for the determined achievements and with the help of a specialist lecturer in the field of Turkish education, a performance task of "Look at a picture and write a story based on this picture" where the writing ability can be best put out. To perform the class activity part of the performance task, 104 students were asked to write their story by saying "Look at the painting in your hand and use your imagination to write a story that tells you the image you see on this painting".

In order to choose the story to be used in the research, classifications of the stories according to 3 categories determined as "good", "medium", "bad" were asked from 4 teachers, 2 of them being Turkish and 2 of being class teachers. As a result, two stories were selected for each category, and it was decided to carry out the research using these 6 stories. Three different scoring keys were created to rate the 6 selected stories. These scoring keys are, a checklist, rating scale and analytic rubric. The 6 selected stories were rated by 45 raters with the help of 3 different tools.

In order to analyze the data, a sampling study by bootsrap was performed to conduct the exchange of number of raters and of scoring keys. 100 samples consisting of 2, 3, 5 and 10 raters were selected from 45 raters in the R program and bootstrap sampling. Analyzes for the 400 selected samples were made separately for the checklist, rating scale and analytic rubric in G theory. In the study, the fully crossed pattern t x sk x r (t: task, sk: scoring key, r: rater) was used. By using this pattern, the reliability of the scales based on these three different measuring instruments was calculated by calculating 2, 3, 5 and 10 raters G coefficients. Since the reliability coefficients obtained in each case are not affected by extreme values, median and standard errors are calculated because the measurements will describe the point where they accumulate better (Turgut and Baykul, 2012; Arıcı, 2005).

The aim of this research is to examine the rater reliability in the context of Generalizability Theory when the same performance tasks are rated by different numbers of raters with the help of a checklist, rating scale and analytical rubric. To this end, answers to the following questions were sought.

1) According to generalizability theory, what are the G coefficient values of the same performance task estimated via generalizability theory of the 100 samples with 2, 3, 5, and 10 raters, selected from the 45 raters and by using the checklist, rating scale and analytic rubric?

2) According to generalizability theory, what are median and standart error of the scoring relability of the same performance task estimated via generalizability theory of the 100 samples with 2, 3, 5, and 10 raters, selected from the 45 raters and by using the checklist, rating scale and analytic rubric?

**Findings (Results)**

The reliability coefficients obtained for the first and second research questions are given in Table 1, Table 2, Table 3, Table 4 and Table 5. When examined from the point of view of 2 rater samples, it is seen that when the scoring is done using the checklist (2 categories), G theory of reliability estimates median is 0,309 and standard errors are 0.024; when analytical rubrics are rated (3 categories), the reliability of G-theory estimates median is 0,439, and the standard errors are 0.024; and when the scoring is done using the rating scale (5 categories), the reliability of G-theory estimates median is 0,474 and the standard errors are 0,030.

When examined from the point of view of 3 rater samples, it is seen that when the scoring is done using the checklist (2 categories), G theory of reliability estimates median is 0,234 and standard errors are 0.020; when

analytical rubrics are rated (3 categories), the reliability of G-theory estimates median is 0,515, and the standard errors are 0.024; and when the scoring is done using the rating scale (5 categories), the reliability of G-theory estimates median is 0,473 and the standard errors are 0,024.

When examined from the point of view of 5 rater samples, it is seen that when the rating is done using the checklist (2 categories), G theory of reliability estimates median is 0,830 and standard errors are 0.020; when analytical rubrics are rated (3 categories), the reliability of G-theory estimates median is 0,589, and the standard errors are 0.019; and when the scoring is done using the rating scale (5 categories), the reliability of G-theory estimates median is 0,605 and the standard errors are 0,020.

When examined from the point of view of 10 rater samples, it is seen that when the rating is done using the checklist (2 categories), G theory of reliability estimates median is 0,363 and standard errors are 0.011; when analytical rubrics are rated (3 categories), the reliability of G-theory estimates median is 0,595, and the standard errors are 0.013; and when the scoring is done using the rating scale (5 categories), the reliability of G-theory estimates median is 0,661 and the standard errors are 0,012.

**Conclusion and Discussion**

When the median values of the reliability estimates of the G theory were examined in this study, it was observed that the median values of 5 raters were increased by the number of raters except for the median of the reliability of the scorings using the checklist and the number of categories at the same time. It has been observed that the standard errors obtained from G theory decreased as the number of raters increased. It has been determined that the lowest standard error values are obtained in the case of 10 raters. When the number of raters is 5 and the number of categories is 2, it can be said that the reliability of G-theory gives the highest value.

Based on the results obtained in this research, researchers can plan to study the reliability of studies between raters based on the theory of G, with two categories and five raters. This study was conducted with 2, 3, 5 and 10 rater samples. Researchers who will conduct similar studies can arrange their research using different evaluator numbers in the light of their decision-making studies. The item counts of the different scoring keys used in the research are fixed but the reliability coefficients can also be examined by using different item counts. Different numbers and types of measuring tools can be used in addition to the three different scoring keys used in this research. In the evaluation of the reliability of this study, the analysis of the G theory is performed with the multivariate crossed t x sk x r pattern. Studies using a nested pattern will provide useful information on the examination of similar studies because using this pattern there may be internal validity problems.

**GİRİŞ**

Eğitimde ölçme ve değerlendirme çalışmalarının önemli bir bölümünü öğrenciler hakkında karar verebilmek amacıyla yapılan çalışmalar oluşturur. Öğrencilerin bir derste başarılı olmak için gerekli olan ön koşul düzeydeki bilgi ve becerilere sahip olup olmadıklarını, bir dersteki öğrenme eksikliklerini, yanlışlıklarını, öğrenme düzeylerini belirlemek ve çeşitli konularda öğrencilere rehberlik çalışmaları yapmak vb. amacıyla ölçme değerlendirme çalışmalarına başvurulur.

Öğrenciler hakkında bilgi toplayabilmek ve çeşitli kararlar alabilmek için farklı ölçme ve değerlendirme araç ve yöntemlerinden yararlanılır. Bu araçlar, uzun cevaplı, kısa cevaplı, doğru yanlış ve çoktan seçmeli maddelerden oluşan ölçme araçları, performans görevleri, projeler, gelişim dosyaları gibi etkinlikler ve kontrol listeleri, dereceleme ölçekleri, dereceli puanlama anahtarları şeklinde sıralanabilir.

Her ölçme aracı farklı durumlarda, farklı düzeydeki kazanımlarda diğerlerine göre daha uygun olabilir. Ölçme aracı hazırlayacak olan eğitimcilerin hangi araçtan yararlanacaklarına öncelikle amaca ve ölçülecek özelliğe bağlı olmak üzere, öğrenci özellikleri, süre vb. ölçütleri dikkate alarak karar vermeleri beklenir. Örneğin performans ölçümlerinde genellikle kontrol listeleri, dereceleme ölçekleri ve rubrikler kullanılmaktadır.

Performans, bireyin sahip olduğu kapasiteyi belli zaman dilimi içinde bir işi başarıyla tamamlamak için kullanabilme yüzdesi olarak tanımlanabilir (Deliceoğlu, 2009). Performans değerlendirme ise bireylerin aktif öğrenme yoluyla süreç içerisinde gerçekleştirdikleri çalışmaların ve süreç sonunda ortaya çıkardıkları ürünlerin değerlendirilmesi şeklinde ifade edilebilir (Tekindal, 2008).

Öğrenci performansının gözlenmesinde yararlanılabilecek etkinliklerden biri performans görevleridir. Performans görevleri öğrencilerin bilgiyi toplama, düzenleme, analiz etme ve yorumlamalarını içeren "araştırmaya dayalı-genişletilmiş cevaplı" etkinlikler olarak gerçekleştirilebileceği gibi, sınıf içerisinde öğretmen kontrolü altında yapılabilecek "sınırlandırılmış" etkinlikler olarak da yapılabilir. Her iki etkinlik çalışmasında da öğrencilerin gerçek yaşamda karşılaşabilecekleri problem durumları sunulmakta ve üst düzey düşünme becerilerinin geliştirilmesi amaçlanmaktadır (Kutlu, Doğan, Karakaya, 2008: s28).

Kontrol listeleri, dereceleme ölçekleri ve dereceli puanlama anahtarları (holistik rubrik ve analitik rubrik) performans görevinin değerlendirilmesinde puanlama anahtarı olarak kullanılabilecek araçlardır. Kontrol listeleri, içinde bir dizi uyarıcının (kazanım, soru, ölçüt) yer aldığı ve puanlayıcının (cevaplayıcının) bu uyarıcılara var/yok, evet/hayır vb. şeklinde verilen ikili kategorilerden birini seçerek tepkide bulunduğu ölçme araçlarıdır. Diğer ifadeyle kontrol listeleri, gösterilen performansın ya da değerlendirilen kişinin sahip olması gereken özelliklere sahip olup olmadığını ortaya koyar (Aiken, 2000). Ancak kontrol listeleri, performansın düzeylerinin ya da ölçütlerin ne dereceye kadar karşılandığı hakkında bilgi vermez (Moskal ve Leydens, 2000). Kontrol listeleri özdeğerlendirme, akran değerlendirme ve grup değerlendirme gibi çalışmalarda da yaygın olarak kullanılmaktadır. Bu araç, eğitim, sanayi, tıp gibi birçok alanda yaygın olarak kullanılan (Aiken, 2000), istenilen bilgiye çabuk ulaşılması, kolay uygulanması, sonuçların daha anlaşılabilir olması açısından avantajlı bir ölçme aracıdır (Hobart ve Frankel, 1999).

Eğitim sisteminde ölçme ve/veya değerlendirme amacıyla yararlanılan bir diğer araç, dereceleme ölçekleridir. Dereceleme ölçekleri hem ilgilenilen özelliğin bireyde bulunma düzeyini ortaya koymak amacıyla ölçme aracı

olarak kullanılabilirken hem de performans görevlerinin puanlanmasında bir puanlama anahtarı olarak kullanılabilir.

Dereceleme ölçekleri, birey ya da nesnelerin belli bir özelliğini doğrudan ölçmek için bir araç bulunmuyorsa, her bir özelliğin bulunuş derecesini belirlemek için kullanılır (Keeves, 1988). Bu araçlar, ölçülen özelliğe ilişkin performans ölçütlerinin ne dereceye kadar karşılandığını görebilmek açısından uygundur (Moskal ve Leydens, 2000). Bir başka deyişle kontrol listeleri yalnızca ilgilenilen özelliğin bireyde bulunup bulunmadığını ortaya koyarken, dereceleme ölçekleri bu özelliklerin yalnızca bireyde bulunup bulunmadığıyla değil, aynı zamanda bireyde bulunma düzeyi/sıklığı ile de ilgilenmektedir. Bu durum, dereceleme ölçeklerinin en az üç kategorili olarak (evet/kısmen/hayır gibi) düzenlenebileceğini de ortaya koymaktadır (Tekindal, 2008).

Performans görevlerini puanlamanın bir başka yolu, dereceli puanlama anahtarından yararlanmaktır. Dereceli puanlama anahtarı, holistik (bütüncül) ve analitik (ayrıntılı) puanlama anahtarı olmak üzere iki grupta incelenebilir. Dereceli puanlama anahtarları (rubrikler); değerlendirme ölçütleri, ölçüt tanımlamaları ve puanlama stratejisi olmak üzere üç bölümden oluşur (Popham, 1997). Değerlendirme ölçütleri; kabul edilebilir yanıtları kabul edilemez yanıtlardan ayırmak için kullanılır. Ölçüt tanımlamaları; öğrencilerin değerlendirilmek istenen yanıtlarındaki niteliksel farklılıkları tanımlama yolunu ifade eder. Puanlama stratejisi; bütünsel (holistic) ya da analitik (analitical) puanlama biçiminde olabilir. Popham'a (1997) göre, bütünsel (holistic) puanlama, performansın daha küçük parçalara ayrılamadığı durumlarda kullanılan ve değerlendirmenin genel olarak yapıldığı durumlarda kullanılırken; analitik (analitical) puanlama, sergilenen performansın en küçük alt parçalarının ayrı ayrı puanlandığı durumlarda kullanılır. Dereceli puanlama anahtarlarından hangisinin kullanılacağına değerlendirmenin amacına bağlı olarak karar verilir.

Eğitimde ve psikolojide bireyler hakkında verilen kararların diğer ifadeyle yapılan değerlendirmelerin isabetliliği, bu kararları almada kullanılan ölçme sonuçlarının niteliği, bu sonuçları elde etmemize olanak sağlayan ölçme araçlarının niteliğine bağlıdır. Bir ölçme aracının seçkisiz hatalardan arınık ölçüm yapabilme yeterliliği olarak adlandırılan güvenirlik ile kullanılma amacına hizmet etme derecesi olarak adlandırılan geçerlik, ölçme araçlarında bulunması gereken iki önemli psikometrik özelliktir (Turgut ve Baykul, 2010).

Performans görevlerine ilişkin ölçmelerde güvenirlik çalışması yapılırken, puanlayıcı güvenirliğinin (intrarater reliability) ve/veya puanlayıcılar arası güvenirliğin (interrater reliability) incelenmesine de ihtiyaç duyulabilir. Puanlayıcı güvenirliği, bir puanlayıcının farklı zamanlarda aynı bireylerin kâğıdını birden fazla puanlamasından elde edilen verilerin tutarlığı olarak ifade edilebilir. Puanlayıcılar arası güvenirlik ise, iki ya da daha çok puanlayıcının aynı bireyi birbirlerinden bağımsız olarak puanlamalarından elde edilen puanların tutarlılık derecesi olarak tanımlanır (Anastasi ve Urbina, 1997). G kuramı (GK) yoluyla puanlayıcı güvenirliği üzerinde çalışmalar yapılabilir.

Genellenebilirlik kuramı (G Kuramı), davranış ölçümlerinde güvenirliğin değerlendirilmesini, güvenilir gözlemlerin tasarımını, araştırılmasını ve kavramlaştırılmasını sağlayan bir istatistiksel kuramdır (Brennan, 2001). G kuramının temeli, varyans analizi (ANOVA) üzerine kurulmuştur. Varyans analiziyle, toplam varyans desendeki varyans bileşenlerine bölünür. Böylece ölçme sonuçları farklı varyans kaynaklarına ayrılarak bireylerin ya da objelerin gözlenen puanlarının evren puanlarına (gerçek puanlarına) genellenebilmesi sağlanır

(Brennan, 2001). Bu varyans kaynakları G kuramında değişkenlik kaynağı (facet) olarak adlandırılır. Bu değişkenlik kaynaklarının da koşulları (conditions) ya da seviyeleri (levels) vardır. Değişkenlik kaynağı ve koşul ifadeleri varyans analizi alanyazınında faktör ve düzey kavramlarına karşılık gelmektedir (Brennan, 2001; Atılgan, 2004).

G kuramı'na göre değişkenlik kaynakları çapraz (crossed) ya da yuvalanmış (nested) şekilde olabilir. Bir değişkenlik kaynağının koşulları başka bir değişkenlik kaynağının koşullarıyla örtüştüğü duruma çapraz desen denilmektedir (Brennan 2001). Çapraz olarak tasarlanmış bir desende değişkenlik kaynakları arasında "x" işareti kullanılır. Örneğin; öğrenciler aynı ürünler doğrultusunda aynı puanlayıcılar tarafından puanlanmış ise öğrenci, ürün ve puanlayıcı değişkenleri çaprazlanarak ö x ü x p desenini oluşturur (ö: öğrenci, ü: ürün, p: puanlayıcı). Değişkenlik kaynaklarının diğer bazı değişkenlik kaynaklarıyla örtüşmediği yani çaprazlanmadığı durumlar da söz konusudur. Bir değişkenlik kaynağının koşulları diğer değişkenlik kaynaklarının bazı koşullarıyla örtüşüyorsa değişkenlik kaynakları yuvalanmış olarak tasarlanmalıdır. Yuvalanmış olarak tasarlanmış bir desende değişkenler arasında ":" işareti kullanılır (Brennan 2001).

G kuramı'nda güvenirliğin araştırılmasında (1) Genellenebilirlik (G) çalışması ve (2) Karar (K) çalışması olmak üzere iki temel çalışma yer almaktadır (Brennan, 2001; Güler, 2011). G çalışması ile varyansların bağıl büyüklüklerine dayalı olarak ana ve ortak etkilerin yorumlanması ve aynı zamanda güvenirlik katsayılarının (G ve Phi) hesaplanması için gerekli değerler sağlanmaktadır. K çalışmasıyla ise madde ve/veya puanlayıcı sayılarının değişimlenmesi yoluyla en güvenilir ve en verimli ölçme desenleri belirlenmektedir (Atılgan, 2005).

Bu araştırmanın amacı, aynı performans görevlerinin farklı sayıda puanlayıcı tarafından kontrol listesi, dereceleme ölçeği ve analitik rubrik yardımıyla puanlanması durumunda, puanlayıcı güvenirliklerinin G kuramı çerçevesinde incelenmesidir. Alanyazında analitik ve holistik rubrik kullanılarak puanlayıcı güvenirliğinin G kuramı ile incelendiği (Büyükkıdık ve Anıl, 2015), performansın dereceli puanlama anahtarı ile puanlanmasından elde edilen puanların güvenirliklerinin G kuramı ile incelendiği (Çakıcı Eser ve Gelbal, 2013), dereceleme ölçeği kullanılarak G kuramı ve Klasik Test Kuramı ile güvenirliklerin incelendiği (Deliceoğlu, 2009; Deliceoğlu ve Çıkrıkçı Demirtaşlı, 2012), puanlanan performansların güvenirliklerinin G kuramı ile incelendiği (Nalbantoğlu Yılmaz ve Başusta, 2015; Nalbantoğlu Yılmaz ve Gelbal, 2011), öğrencilerin yazma becerilerinin analitik ve holistik rubrik ile puanlanarak puanlayıcı güvenirliğinin incelendiği (Covill, 2012), holistik ve analitik rubrik ile yapılan değerlendirmelerin karşılaştırıldığı (Singer ve LeMahieu, 2011) çalışmalar mevcuttur. Fakat yapılan alanyazın araştırmasında, performans görevlerinin farklı sayıda puanlayıcılarla ve/veya üç farklı ölçme aracı yardımıyla puanlanmasına ilişkin bir çalışmaya rastlanmamıştır. Bu anlamda, araştırma bulgularının ölçme değerlendirme alanyazınına önemli bir katkı sağlaması beklenmektedir.

Bu araştırma kapsamında iki araştırma sorusuna yanıt aranmıştır. (I) G kuramı'na göre, aynı performans görevlerinin; kontrol listesi, dereceleme ölçeği, analitik rubrik kullanılarak ve 2, 3, 5, 10 puanlayıcıdan oluşan 100'er örneklemin puanlama güvenirlikleri nasıldır? (II) G kuramı'na göre, aynı performans görevlerinin; kontrol listesi, dereceleme ölçeği, analitik rubrik kullanılarak ve 2, 3, 5, 10 puanlayıcıdan oluşan 100'er örneklemin puanlama güvenirliklerinin ortancaları ve standart hataları nasıldır?

**YÖNTEM**

Bu çalışmanın genel amacı performans değerlendirmelerinin güvenirliğinin incelenmesinde farklı ölçme araçlarının (kontrol listesi, dereceleme ölçeği, analitik rubrik) kullanıldığı ve puanlayıcı sayısının değişimlendiği koşullarda G kuramından elde edilen sonuçları karşılaştırmalı olarak incelemektir. Bu yönüyle çalışma G kuramına yönelik bilgi üretmeyi amaçladığından, temel araştırma niteliğindedir (Singh, 2006; Kothari, 2004).

**Katılımcılar**

Çalışma 5. sınıfta okuyan 104 öğrenci ve çeşitli okullarda görev yapan 104 öğretmen üzerinden yürütülmüştür. Öğrenciler kendilerine verilen performans görevini yerine getiren grubu; öğretmenler ise bu performans görevlerini değerlendiren puanlayıcı grubunu oluşturmaktadır. Çalışma sürecinde puanlayıcı grubunda yaşanan katılımcı kayıpları nedeniyle araştırma, 45 puanlayıcıdan (öğretmenden) elde edilen veriler üzerinden gerçekleştirilmiştir. 45 puanlayıcının %62,16'sını kadınlar, %37,74'sini erkekler oluşturmaktadır. 21 yıl ve üstünde görev yapanlar grubun %51,06'sını, 11-20 yıl arası görev yapanlar grubun %42,2'sini ve 1-10 yıl görev yapanlar grubun %6,67'sini oluşturmaktadır.

**Performans Görevinin Hazırlanması ve Uygulanması**

Bir kontrol listesi hazırlanarak 5. Sınıf Türkçe dersi yazma becerileri listelenerek 7 sınıf öğretmenine bu kazanımların hangilerinin performans görevi yardımıyla gözlemlenebileceğine karar vermeleri istenmiştir. Basit uyum indeksi kullanılarak %70 (Erkuş, 2006) ve üzeri uyumun sağlandığı kazanımlar belirlenmiştir. Belirlenen kazanımlara yönelik öğretmenlerden performans görevi örnekleri istenmiş ve Türkçe eğitimi alanında uzman bir öğretim üyesi yardımıyla yazma becerisinin en iyi ortaya koyulabileceği "Bir resme bakar ve bu resmi temel alarak bir hikâye yazar" performans görevi seçilmiştir. Performans görevinin sınıf dışı etkinlik kısmını gerçekleştirmek üzere onlardan 5'er kısa hikâye okumaları ve bunlardan bir tanesinin özetini yazmaları istenmiştir. Daha sonra araştırmacı performans görevinin sınıf içi etkinlik kısmını gerçekleştirmek üzere sınıflara gitmiştir. Öğrencilere bir tablo göstererek "elinizdeki tabloya bakın ve hayal gücünüzü kullanarak bu tabloda gördüğünüz resmi anlatan bir hikâye yazın" diyerek 104 öğrenciden hikayelerini yazmalarını istemiştir.

104 öğrenci tarafından yazılan hikâyeler, araştırmacılar tarafından incelenmiş; anlaşılamayan, okunamayan, performans görevinin amacının dışında hazırlanan hikâyeler elenmiştir. Bu eleme sonucunda toplam hikâye sayısı 54'e düşmüştür. 54 hikâye arasından, araştırmada kullanılmak üzere hikâye seçmek amacıyla, 2'si Türkçe, 2'si Sınıf öğretmeni olmak üzere 4 öğretmenden "iyi", "orta", "kötü" olmak üzere belirlenen 3 kategorilere göre hikâyeleri sınıflamaları istenmiştir. 4 öğretmenin de ortaklaştığı yani aynı hikâyeyi aynı kategoriye yerleştirdiği 17 hikâye olduğu tespit edilmiştir. Bunun sonucunda her kategoriden seçkisiz olarak 2'şer hikâye seçilmiş ve araştırmanın bu 6 hikâye kullanılarak yürütülmesi kararlaştırılmıştır.

**Veri Toplama Araçları**

Seçilen 6 hikayenin puanlanması amacıyla üç farklı puanlama anahtarı oluşturulmuştur. Bu puanlama anahtarları, kontrol listesi dereceleme ölçeği ve analitik rubriktir. Araçların geliştirilmesi aşamasında öncelikle Türkçe Eğitimi alanında bir öğretim üyesi (Doçent doktor), yine Türkçe Eğitimi alanında uzman olan bir araştırma görevlisi ve 8 Türkçe öğretmeniyle görüşülerek onlara yazılan hikâyeyi değerlendirmede göz önüne alınması gereken ölçütlerin (davranış göstergelerinin) neler olduğu sorulmuş ve ilgili alan yazın taraması da gerçekleştirilerek bu ölçütler listelenmiştir.

Listelenen ölçütler ana ve alt ölçütler başlıkları altına yerleştirilmiş ve davranış göstergelerinin ana ve alt ölçütlere uygunluğunun ve puanlama yapılırken ağırlıklandırmanın nasıl olması gerektiğinin kararlaştırılabilmesi amacıyla bir değerlendirme formu hazırlanmıştır 9 Türkçe öğretmeni bu formu dikkate alarak listelenen ölçüt ve alt ölçütlerin uygun olup olmadığına ve bu ölçüt ve alt ölçütlerin puan ağırlıklarına karar vermişlerdir. 9 öğretmenin her ölçüt/alt ölçüt için yaptığı değerlendirmeler arasındaki uyum basit uyum katsayısı ile incelenmiştir. %70 ve üzerinde uyum gösteren ölçütler puanlama anahtarlarına dâhil edilmiştir (Erkuş, 2006).

İlgili kazanımlar performans göstergesi olarak kabul edilmiş ve aynı kazanımlar için "Evet-Hayır" kategorilerinden oluşan bir kontrol listesi, "0-4" derecelerinden oluşan bir dereceleme ölçeği ve "1-3" derecelerinden oluşan ve her bir derece için performans tanımlarının yapıldığı bir analitik rubrik hazırlanmıştır. Temel amaç, puanlama işlemi ayrıntılandırıldıkça puanlayıcı güvenirliğinin nasıl değiştiğini ya da puanlayıcı güvenirliğindeki değişimi gözlemlemek olduğundan kontrol listesi ve dereceleme ölçeği gözetilerek Holistik yerine Analitik rubrik kullanımı kararlaştırılmıştır.

Seçilen 6 hikâye, 45 puanlayıcı tarafından, hazırlanan 3 farklı araç yardımıyla puanlanmıştır. Sıra etkisini ortadan kaldırmak amacıyla araçlar farklı puanlayıcılara farklı sırayla ve 10-15 gün arayla verilmiştir. Sonuç olarak puanlayıcılar yaklaşık 50 günlük bir süre içinde 6 hikâyeyi 3 farklı araç yardımıyla derecelendirmiştir.

**Verilerin Analizi**

Verilerin analiz edilebilmesi için öncelikle, 6 hikâye 45 puanlayıcı ve 3 farklı araç yardımıyla elde edilen verilerden, puanlayıcı sayısının ve puanlama anahtarlarının değişimlenmesi çalışmasını gerçekleştirmek üzere bootsrap yoluyla örnekleme çalışması gerçekleştirilmiştir. Efron ve Tibshirani (1993), evrenden seçilen seçkisiz bir örneklemin verilerinin yerlerini değiştirmeyle, seçkisiz ve farklı birçok örneklem çekilerek ve çekilen her örneklem için ilgilenilen istatistikler hesaplandığı "yeniden örnekleme (bootstrap)" yöntemini önermiştir. Elde edilen değerlerin standart sapmaları hesaplandığında yapılan kestirime ilişkin standart hata değeri elde edilmektedir (Akt:Kolen ve Brennan, 2004). Efron ve Tibshirani (1993) çekilecek örneklem sayısının 25 ile 200 arasında olmasının uygun olacağını belirtmişlerdir (Akt:Kolen ve Brennan, 2004). Alanda G kuramı ile ilgili analizlerin yapılabileceği çeşitli paket programlar mevcut olmakla birlikte (EduG, G-String vb.) Bu çalışmanın kolaylığı açısından R programı yardımıyla örnekleme işlemi hesaplamalar gerçekleştirilmiştir. R programında

yazılan kodlar ve bootstrap örnekleme sonucunda 45 puanlayıcı arasından seçkisiz olarak 2, 3, 5 ve 10 puanlayıcıdan oluşan 100'er örneklem seçilmiştir. Örneğin, 2 puanlayıcı kombinasyonundan oluşan 100 farklı örneklem; 3 puanlayıcı kombinasyonundan oluşan 100 farklı örneklem vb. elde edilmiştir. Seçilen 400 farklı örneklem için analizler G kuramı'nda kontrol listesi, dereceleme ölçeği ve analitik rubrik için ayrı ayrı yapılmıştır. Araştırmada tümüyle çaprazlanmış desen g x pa x p (g: görev, pa: puanlama anahtarı, p: puanlayıcı) kullanılmıştır. Bu desen kullanılarak, 2, 3, 5 ve 10 puanlayıcı olması durumunda bu üç farklı ölçme aracına bağlı olarak yapılan puanlamaların güvenirliği G katsayısı hesaplanarak incelenmiştir. G kuramı'nda 2, 3, 5 ve 10 puanlayıcılı örneklemlerde yapılacak olan güvenirlik kestirimi için birer kod yazılmıştır. 2, 3, 5 ve 10 puanlayıcılar için ayrı ayrı olmak üzere yazılan kodlar birleştirilerek 1200 örneklemin her biri için güvenirlik katsayıları hesaplanmıştır. Örneklemlerin seçilmesi, gerekli düzenlemelerin yapılması güvenirliklerin hesaplanmasında R programından yararlanılmıştır (yazılan kodlara Aktaş'ın (2013) yüksek lisans tezinden ulaşılabilir). Elde edilen güvenirlik katsayılarının her durum için 100'er örneklemlerinin, uç değerlerden etkilenmediğinden ölçümlerin yığıldıkları noktayı daha iyi betimleyeceği için (Turgut ve Baykul, 2012; Arıcı, 2005) ortancaları ve standart hataları hesaplanmıştır.

Bu araştırmanın amacı, aynı performans görevlerinin farklı sayıda puanlayıcı tarafından kontrol listesi, dereceleme ölçeği ve analitik rubrik yardımıyla puanlanması durumunda, puanlayıcı güvenirliklerinin Genellenebilirlik Kuramı çerçevesinde incelenmesidir. Bu amaç doğrultusunda, aşağıdaki sorulara yanıt aranmıştır.

1) Genellenebilirlik kuramına göre, aynı performans görevlerinin; Kontrol listesi, Dereceleme ölçeği ve Analitik rubrik kullanılarak ve 45 puanlayıcı arasından seçkisiz olarak seçilen, 2'şer, 3'er, 5'er ve 10'ar puanlayıcıdan oluşan 100'er örneklemin genellenebilirlik kuramına göre kestirilen G katsayılarının derecesi nasıldır?

2) Genellenebilirlik kuramına göre, aynı performans görevlerinin; Kontrol listesi, Dereceleme ölçeği ve Analitik rubrik kullanılarak ve 45 puanlayıcı arasından seçkisiz olarak seçilen, 2'şer, 3'er, 5'er ve 10'ar puanlayıcıdan oluşan 100'er örneklemin puanlama güvenirliklerinin ortancaları ve standart hataları nasıldır?

**BULGULAR**

**I. Araştırma Sorusuna İlişkin Bulgular**

Araştırmanın ilk sorusuna yanıt aramak amacıyla, 2, 3, 5 ve 10 puanlayıcı için ayrı ayrı analizler gerçekleştirilmiştir. 2 puanlayıcıdan oluşan 100 örneklem için, 3 farklı araç kullanılarak yapılan puanlamaların güvenirliklerine ilişkin bulgular Tablo 1'de verilmiştir.

**Tablo1.** Üç Farklı Puanlama Anahtarı Kullanılarak ve 2 Puanlayıcıdan Oluşan 100'er Örneklemin Puanlama Güvenirlikleri

| No | KL | DÖ | AR | No | KL | DÖ | AR | No | KL | DÖ | AR |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0,000 | 0,000 | 0,000 | 36 | 0,170 | 0,347 | 0,275 | 71 | 0,433 | 0,554 | 0,608 |
| 2 | 0,000 | 0,000 | 0,000 | 37 | 0,170 | 0,348 | 0,280 | 72 | 0,435 | 0,567 | 0,635 |
| 3 | 0,000 | 0,000 | 0,000 | 38 | 0,202 | 0,355 | 0,311 | 73 | 0,435 | 0,577 | 0,641 |
| 4 | 0,000 | 0,000 | 0,000 | 39 | 0,203 | 0,355 | 0,311 | 74 | 0,454 | 0,578 | 0,659 |
| 5 | 0,000 | 0,000 | 0,000 | 40 | 0,203 | 0,356 | 0,327 | 75 | 0,456 | 0,588 | 0,680 |
| 6 | 0,000 | 0,000 | 0,000 | 41 | 0,209 | 0,360 | 0,355 | 76 | 0,460 | 0,600 | 0,685 |
| 7 | 0,000 | 0,000 | 0,000 | 42 | 0,219 | 0,364 | 0,388 | 77 | 0,464 | 0,612 | 0,690 |
| 8 | 0,000 | 0,000 | 0,000 | 43 | 0,238 | 0,378 | 0,389 | 78 | 0,473 | 0,621 | 0,696 |
| 9 | 0,000 | 0,000 | 0,000 | 44 | 0,252 | 0,407 | 0,406 | 79 | 0,474 | 0,634 | 0,698 |
| 10 | 0,000 | 0,000 | 0,000 | 45 | 0,270 | 0,423 | 0,410 | 80 | 0,490 | 0,665 | 0,757 |
| 11 | 0,000 | 0,000 | 0,000 | 46 | 0,276 | 0,423 | 0,417 | 81 | 0,506 | 0,667 | 0,759 |
| 12 | 0,000 | 0,000 | 0,000 | 47 | 0,300 | 0,430 | 0,430 | 82 | 0,516 | 0,675 | 0,766 |
| 13 | 0,000 | 0,000 | 0,000 | 48 | 0,300 | 0,431 | 0,461 | 83 | 0,529 | 0,675 | 0,772 |
| 14 | 0,000 | 0,000 | 0,000 | 49 | 0,302 | 0,431 | 0,468 | 84 | 0,529 | 0,677 | 0,773 |
| 15 | 0,000 | 0,000 | 0,000 | 50 | 0,302 | 0,438 | 0,468 | 85 | 0,540 | 0,682 | 0,787 |
| 16 | 0,000 | 0,032 | 0,014 | 51 | 0,315 | 0,440 | 0,479 | 86 | 0,556 | 0,683 | 0,788 |
| 17 | 0,000 | 0,063 | 0,022 | 52 | 0,315 | 0,442 | 0,479 | 87 | 0,599 | 0,690 | 0,788 |
| 18 | 0,000 | 0,134 | 0,022 | 53 | 0,330 | 0,446 | 0,483 | 88 | 0,599 | 0,701 | 0,790 |
| 19 | 0,000 | 0,207 | 0,078 | 54 | 0,339 | 0,446 | 0,484 | 89 | 0,630 | 0,704 | 0,809 |
| 20 | 0,000 | 0,216 | 0,078 | 55 | 0,347 | 0,446 | 0,490 | 90 | 0,643 | 0,719 | 0,815 |
| 21 | 0,000 | 0,217 | 0,095 | 56 | 0,352 | 0,473 | 0,499 | 91 | 0,675 | 0,728 | 0,827 |
| 22 | 0,003 | 0,240 | 0,108 | 57 | 0,364 | 0,473 | 0,511 | 92 | 0,684 | 0,737 | 0,829 |
| 23 | 0,006 | 0,247 | 0,129 | 58 | 0,367 | 0,476 | 0,514 | 93 | 0,696 | 0,748 | 0,841 |
| 24 | 0,013 | 0,266 | 0,139 | 59 | 0,382 | 0,480 | 0,531 | 94 | 0,717 | 0,753 | 0,883 |
| 25 | 0,015 | 0,266 | 0,144 | 60 | 0,382 | 0,487 | 0,533 | 95 | 0,727 | 0,772 | 0,884 |
| 26 | 0,054 | 0,288 | 0,147 | 61 | 0,385 | 0,487 | 0,537 | 96 | 0,740 | 0,772 | 0,890 |
| 27 | 0,056 | 0,298 | 0,148 | 62 | 0,385 | 0,491 | 0,543 | 97 | 0,751 | 0,780 | 0,899 |
| 28 | 0,056 | 0,300 | 0,163 | 63 | 0,391 | 0,498 | 0,546 | 98 | 0,788 | 0,794 | 0,930 |
| 29 | 0,069 | 0,309 | 0,191 | 64 | 0,391 | 0,506 | 0,553 | 99 | 0,796 | 0,797 | 0,950 |
| 30 | 0,084 | 0,311 | 0,232 | 65 | 0,394 | 0,508 | 0,569 | 100 | 0,800 | 0,880 | 0,950 |
| 31 | 0,092 | 0,324 | 0,237 | 66 | 0,401 | 0,517 | 0,577 | | | | |
| 32 | 0,122 | 0,334 | 0,238 | 67 | 0,403 | 0,536 | 0,585 | | | | |
| 33 | 0,153 | 0,343 | 0,259 | 68 | 0,406 | 0,536 | 0,599 | | | | |
| 34 | 0,154 | 0,343 | 0,268 | 69 | 0,429 | 0,541 | 0,600 | | | | |
| 35 | 0,162 | 0,345 | 0,268 | 70 | 0,431 | 0,553 | 0,605 | | | | |

KL: kontrol listesi, DÖ: dereceli puanlama anahtarı, AR: analitik rubrik

Tablo 1, incelendiğinde, 2 puanlayıcılı örneklemlerde kontrol listesi kullanılarak puanlama yapıldığında (2 kategorili) güvenirlik kestirimleri 0,000 ile 0,800 arasında; analitik rubrik kullanılarak puanlama yapıldığında (3 kategorili) güvenirlik kestirimlerinin 0,000 ile 0,880 arasında; dereceleme ölçeği kullanılarak puanlama yapıldığında (5 kategorili) güvenirlik kestirimlerinin 0,000 ile 0,950 arasında değiştiği gözlenmektedir.

3 puanlayıcıdan oluşan 100 örneklem için, 3 farklı araç kullanılarak yapılan puanlamaların güvenirliklerine ilişkin bulgular Tablo 2'de verilmiştir.

**Tablo 2.** Üç Farklı Puanlama Anahtarı Kullanılarak ve 3 Puanlayıcıdan Oluşan 100'er Örneklemin Puanlama Güvenirlikleri

| No | KL | DÖ | AR | No | KL | DÖ | AR | No | KL | DÖ | AR |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0,000 | 0,000 | 0,000 | 36 | 0,155 | 0,418 | 0,380 | 71 | 0,340 | 0,617 | 0,582 |
| 2 | 0,000 | 0,000 | 0,000 | 37 | 0,166 | 0,438 | 0,384 | 72 | 0,348 | 0,618 | 0,598 |
| 3 | 0,000 | 0,000 | 0,000 | 38 | 0,171 | 0,441 | 0,389 | 73 | 0,349 | 0,621 | 0,598 |
| 4 | 0,000 | 0,000 | 0,000 | 39 | 0,176 | 0,444 | 0,389 | 74 | 0,377 | 0,624 | 0,613 |
| 5 | 0,000 | 0,000 | 0,000 | 40 | 0,184 | 0,447 | 0,394 | 75 | 0,387 | 0,624 | 0,627 |
| 6 | 0,000 | 0,000 | 0,000 | 41 | 0,188 | 0,449 | 0,395 | 76 | 0,393 | 0,628 | 0,646 |
| 7 | 0,000 | 0,000 | 0,000 | 42 | 0,197 | 0,452 | 0,399 | 77 | 0,394 | 0,630 | 0,649 |
| 8 | 0,000 | 0,000 | 0,000 | 43 | 0,202 | 0,461 | 0,412 | 78 | 0,397 | 0,632 | 0,664 |
| 9 | 0,000 | 0,000 | 0,092 | 44 | 0,202 | 0,463 | 0,432 | 79 | 0,413 | 0,633 | 0,671 |
| 10 | 0,000 | 0,000 | 0,117 | 45 | 0,215 | 0,469 | 0,438 | 80 | 0,425 | 0,639 | 0,684 |
| 11 | 0,000 | 0,000 | 0,142 | 46 | 0,217 | 0,470 | 0,455 | 81 | 0,446 | 0,639 | 0,698 |
| 12 | 0,000 | 0,000 | 0,150 | 47 | 0,217 | 0,498 | 0,457 | 82 | 0,451 | 0,642 | 0,701 |
| 13 | 0,000 | 0,000 | 0,164 | 48 | 0,220 | 0,498 | 0,469 | 83 | 0,455 | 0,644 | 0,710 |
| 14 | 0,000 | 0,000 | 0,164 | 49 | 0,226 | 0,504 | 0,470 | 84 | 0,466 | 0,647 | 0,713 |
| 15 | 0,017 | 0,010 | 0,174 | 50 | 0,227 | 0,513 | 0,472 | 85 | 0,500 | 0,653 | 0,717 |
| 16 | 0,019 | 0,059 | 0,185 | 51 | 0,240 | 0,516 | 0,473 | 86 | 0,505 | 0,655 | 0,720 |
| 17 | 0,020 | 0,093 | 0,200 | 52 | 0,247 | 0,519 | 0,487 | 87 | 0,507 | 0,664 | 0,745 |
| 18 | 0,026 | 0,096 | 0,202 | 53 | 0,252 | 0,520 | 0,492 | 88 | 0,507 | 0,666 | 0,759 |
| 19 | 0,032 | 0,129 | 0,211 | 54 | 0,276 | 0,525 | 0,500 | 89 | 0,511 | 0,667 | 0,764 |
| 20 | 0,037 | 0,133 | 0,241 | 55 | 0,293 | 0,535 | 0,500 | 90 | 0,516 | 0,669 | 0,767 |
| 21 | 0,040 | 0,137 | 0,241 | 56 | 0,294 | 0,536 | 0,502 | 91 | 0,530 | 0,681 | 0,774 |
| 22 | 0,052 | 0,148 | 0,272 | 57 | 0,297 | 0,537 | 0,505 | 92 | 0,553 | 0,698 | 0,776 |
| 23 | 0,054 | 0,191 | 0,286 | 58 | 0,300 | 0,538 | 0,506 | 93 | 0,578 | 0,708 | 0,786 |
| 24 | 0,057 | 0,216 | 0,299 | 59 | 0,302 | 0,549 | 0,506 | 94 | 0,629 | 0,709 | 0,811 |
| 25 | 0,059 | 0,270 | 0,303 | 60 | 0,304 | 0,560 | 0,515 | 95 | 0,646 | 0,740 | 0,836 |
| 26 | 0,059 | 0,348 | 0,325 | 61 | 0,308 | 0,568 | 0,516 | 96 | 0,652 | 0,745 | 0,840 |
| 27 | 0,065 | 0,350 | 0,326 | 62 | 0,309 | 0,571 | 0,525 | 97 | 0,654 | 0,759 | 0,854 |
| 28 | 0,071 | 0,353 | 0,331 | 63 | 0,314 | 0,576 | 0,525 | 98 | 0,696 | 0,776 | 0,887 |
| 29 | 0,099 | 0,356 | 0,336 | 64 | 0,320 | 0,583 | 0,530 | 99 | 0,703 | 0,779 | 0,894 |
| 30 | 0,103 | 0,360 | 0,336 | 65 | 0,320 | 0,586 | 0,531 | 100 | 0,869 | 0,787 | 0,927 |
| 31 | 0,114 | 0,372 | 0,340 | 66 | 0,323 | 0,593 | 0,541 | | | | |
| 32 | 0,116 | 0,382 | 0,341 | 67 | 0,327 | 0,600 | 0,554 | | | | |
| 33 | 0,134 | 0,383 | 0,361 | 68 | 0,332 | 0,601 | 0,565 | | | | |
| 34 | 0,136 | 0,386 | 0,366 | 69 | 0,332 | 0,609 | 0,568 | | | | |
| 35 | 0,138 | 0,399 | 0,373 | 70 | 0,333 | 0,610 | 0,576 | | | | |

Tablo 2, incelendiğinde, 3 puanlayıcılı örneklemlerde kontrol listesi kullanılarak puanlama yapıldığında (2 kategorili) güvenirlik kestirimleri 0,000 ile 0,869 arasında; analitik rubrik kullanılarak puanlama yapıldığında (3 kategorili) güvenirlik kestirimlerinin 0,000 ile 0,787 arasında; dereceleme ölçeği kullanılarak puanlama yapıldığında (5 kategorili) güvenirlik kestirimlerinin 0,000 ile 0,927 arasında değiştiği gözlenmektedir.

5 puanlayıcıdan oluşan 100 örneklem için, 3 farklı araç kullanılarak yapılan puanlamaların güvenirliklerine ilişkin bulgular Tablo 3'te verilmiştir.

**Tablo 3.** Üç Farklı Puanlama Anahtarı Kullanılarak ve 5 Puanlayıcıdan Oluşan 100'er Örneklemin Puanlama Güvenirlikleri

| No | KL | DÖ | AR | No | KL | DÖ | AR | No | KL | DÖ | AR |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0,000 | 0,000 | 0,000 | 36 | 0,752 | 0,524 | 0,525 | 71 | 0,888 | 0,655 | 0,680 |
| 2 | 0,000 | 0,076 | 0,000 | 37 | 0,760 | 0,526 | 0,527 | 72 | 0,888 | 0,660 | 0,681 |
| 3 | 0,084 | 0,118 | 0,024 | 38 | 0,761 | 0,532 | 0,531 | 73 | 0,891 | 0,660 | 0,681 |
| 4 | 0,197 | 0,154 | 0,095 | 39 | 0,773 | 0,537 | 0,533 | 74 | 0,892 | 0,661 | 0,685 |
| 5 | 0,343 | 0,159 | 0,159 | 40 | 0,780 | 0,546 | 0,536 | 75 | 0,894 | 0,662 | 0,693 |
| 6 | 0,387 | 0,178 | 0,179 | 41 | 0,791 | 0,557 | 0,537 | 76 | 0,895 | 0,669 | 0,712 |
| 7 | 0,434 | 0,192 | 0,211 | 42 | 0,800 | 0,557 | 0,546 | 77 | 0,895 | 0,671 | 0,714 |
| 8 | 0,446 | 0,199 | 0,215 | 43 | 0,811 | 0,560 | 0,546 | 78 | 0,897 | 0,672 | 0,715 |
| 9 | 0,451 | 0,201 | 0,217 | 44 | 0,811 | 0,562 | 0,552 | 79 | 0,899 | 0,679 | 0,716 |
| 10 | 0,491 | 0,202 | 0,220 | 45 | 0,817 | 0,569 | 0,556 | 80 | 0,901 | 0,681 | 0,719 |
| 11 | 0,513 | 0,213 | 0,236 | 46 | 0,825 | 0,571 | 0,562 | 81 | 0,902 | 0,685 | 0,719 |
| 12 | 0,520 | 0,218 | 0,247 | 47 | 0,825 | 0,577 | 0,568 | 82 | 0,902 | 0,685 | 0,720 |
| 13 | 0,559 | 0,227 | 0,257 | 48 | 0,826 | 0,580 | 0,581 | 83 | 0,903 | 0,685 | 0,720 |
| 14 | 0,563 | 0,241 | 0,267 | 49 | 0,826 | 0,582 | 0,592 | 84 | 0,907 | 0,691 | 0,727 |
| 15 | 0,577 | 0,308 | 0,269 | 50 | 0,828 | 0,585 | 0,603 | 85 | 0,907 | 0,693 | 0,740 |
| 16 | 0,580 | 0,310 | 0,290 | 51 | 0,832 | 0,592 | 0,607 | 86 | 0,908 | 0,700 | 0,745 |
| 17 | 0,606 | 0,314 | 0,332 | 52 | 0,833 | 0,596 | 0,608 | 87 | 0,908 | 0,705 | 0,753 |
| 18 | 0,606 | 0,353 | 0,333 | 53 | 0,834 | 0,606 | 0,617 | 88 | 0,908 | 0,709 | 0,756 |
| 19 | 0,609 | 0,363 | 0,359 | 54 | 0,838 | 0,609 | 0,626 | 89 | 0,911 | 0,715 | 0,761 |
| 20 | 0,612 | 0,372 | 0,366 | 55 | 0,853 | 0,611 | 0,638 | 90 | 0,913 | 0,720 | 0,764 |
| 21 | 0,638 | 0,373 | 0,371 | 56 | 0,855 | 0,614 | 0,639 | 91 | 0,917 | 0,731 | 0,773 |
| 22 | 0,639 | 0,386 | 0,398 | 57 | 0,858 | 0,618 | 0,639 | 92 | 0,918 | 0,744 | 0,773 |
| 23 | 0,650 | 0,396 | 0,401 | 58 | 0,860 | 0,619 | 0,643 | 93 | 0,919 | 0,745 | 0,780 |
| 24 | 0,667 | 0,404 | 0,424 | 59 | 0,863 | 0,622 | 0,643 | 94 | 0,919 | 0,756 | 0,782 |
| 25 | 0,703 | 0,408 | 0,427 | 60 | 0,864 | 0,623 | 0,645 | 95 | 0,925 | 0,765 | 0,782 |
| 26 | 0,720 | 0,427 | 0,436 | 61 | 0,865 | 0,623 | 0,646 | 96 | 0,930 | 0,766 | 0,801 |
| 27 | 0,723 | 0,441 | 0,458 | 62 | 0,868 | 0,627 | 0,650 | 97 | 0,932 | 0,768 | 0,805 |
| 28 | 0,727 | 0,446 | 0,469 | 63 | 0,871 | 0,629 | 0,656 | 98 | 0,938 | 0,793 | 0,820 |
| 29 | 0,731 | 0,465 | 0,470 | 64 | 0,872 | 0,629 | 0,657 | 99 | 0,940 | 0,793 | 0,860 |
| 30 | 0,735 | 0,470 | 0,483 | 65 | 0,872 | 0,637 | 0,659 | 100 | 0,960 | 0,804 | 0,884 |
| 31 | 0,738 | 0,483 | 0,492 | 66 | 0,874 | 0,637 | 0,660 | | | | |
| 32 | 0,740 | 0,504 | 0,505 | 67 | 0,874 | 0,639 | 0,671 | | | | |
| 33 | 0,747 | 0,509 | 0,515 | 68 | 0,878 | 0,640 | 0,673 | | | | |
| 34 | 0,747 | 0,512 | 0,520 | 69 | 0,878 | 0,644 | 0,677 | | | | |
| 35 | 0,752 | 0,523 | 0,524 | 70 | 0,888 | 0,645 | 0,678 | | | | |

Tablo 3, incelendiğinde, 5 puanlayıcılı örneklemlerde kontrol listesi kullanılarak puanlama yapıldığında (2 kategorili) güvenirlik kestirimleri 0,000 ile 0,960 arasında; analitik rubrik kullanılarak puanlama yapıldığında (3 kategorili) güvenirlik kestirimlerinin 0,000 ile 0,804 arasında; dereceleme ölçeği kullanılarak puanlama yapıldığında (5 kategorili) güvenirlik kestirimlerinin 0,000 ile 0,884 arasında değiştiği gözlenmektedir.

10 puanlayıcıdan oluşan 100 örneklem için, 3 farklı araç kullanılarak yapılan puanlamaların güvenirliklerine ilişkin bulgular Tablo 4'te verilmiştir.

**Tablo 4.** Üç Farklı Puanlama Anahtarı Kullanılarak ve 10 Puanlayıcıdan Oluşan 100'er Örneklemin Puanlama Güvenirlikleri

| No | KL | DÖ | AR | No | KL | DÖ | AR | No | KL | DÖ | AR |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0,041 | 0,214 | 0,278 | 36 | 0,332 | 0,542 | 0,640 | 71 | 0,427 | 0,644 | 0,728 |
| 2 | 0,118 | 0,261 | 0,346 | 37 | 0,335 | 0,548 | 0,642 | 72 | 0,436 | 0,648 | 0,729 |
| 3 | 0,132 | 0,279 | 0,369 | 38 | 0,336 | 0,550 | 0,642 | 73 | 0,436 | 0,664 | 0,736 |
| 4 | 0,161 | 0,282 | 0,412 | 39 | 0,337 | 0,551 | 0,649 | 74 | 0,442 | 0,664 | 0,737 |
| 5 | 0,170 | 0,309 | 0,433 | 40 | 0,338 | 0,556 | 0,650 | 75 | 0,444 | 0,672 | 0,739 |
| 6 | 0,223 | 0,327 | 0,444 | 41 | 0,341 | 0,558 | 0,650 | 76 | 0,447 | 0,676 | 0,743 |
| 7 | 0,233 | 0,338 | 0,449 | 42 | 0,342 | 0,562 | 0,651 | 77 | 0,448 | 0,676 | 0,744 |
| 8 | 0,235 | 0,357 | 0,465 | 43 | 0,344 | 0,572 | 0,651 | 78 | 0,449 | 0,679 | 0,744 |
| 9 | 0,236 | 0,372 | 0,487 | 44 | 0,349 | 0,577 | 0,651 | 79 | 0,452 | 0,681 | 0,753 |
| 10 | 0,245 | 0,375 | 0,490 | 45 | 0,355 | 0,578 | 0,652 | 80 | 0,455 | 0,682 | 0,755 |
| 11 | 0,250 | 0,381 | 0,497 | 46 | 0,356 | 0,580 | 0,653 | 81 | 0,461 | 0,684 | 0,755 |
| 12 | 0,254 | 0,398 | 0,505 | 47 | 0,359 | 0,582 | 0,657 | 82 | 0,462 | 0,685 | 0,757 |
| 13 | 0,262 | 0,399 | 0,529 | 48 | 0,359 | 0,587 | 0,658 | 83 | 0,468 | 0,685 | 0,760 |
| 14 | 0,263 | 0,411 | 0,534 | 49 | 0,361 | 0,588 | 0,659 | 84 | 0,469 | 0,686 | 0,761 |
| 15 | 0,283 | 0,415 | 0,539 | 50 | 0,361 | 0,592 | 0,660 | 85 | 0,477 | 0,690 | 0,762 |
| 16 | 0,285 | 0,415 | 0,540 | 51 | 0,364 | 0,597 | 0,662 | 86 | 0,478 | 0,690 | 0,763 |
| 17 | 0,285 | 0,428 | 0,542 | 52 | 0,365 | 0,600 | 0,663 | 87 | 0,489 | 0,693 | 0,767 |
| 18 | 0,298 | 0,428 | 0,545 | 53 | 0,365 | 0,601 | 0,664 | 88 | 0,491 | 0,697 | 0,768 |
| 19 | 0,301 | 0,433 | 0,551 | 54 | 0,367 | 0,602 | 0,665 | 89 | 0,493 | 0,697 | 0,774 |
| 20 | 0,305 | 0,439 | 0,561 | 55 | 0,376 | 0,603 | 0,666 | 90 | 0,499 | 0,697 | 0,779 |
| 21 | 0,306 | 0,445 | 0,561 | 56 | 0,376 | 0,604 | 0,667 | 91 | 0,500 | 0,701 | 0,791 |
| 22 | 0,307 | 0,445 | 0,570 | 57 | 0,377 | 0,610 | 0,667 | 92 | 0,506 | 0,705 | 0,792 |
| 23 | 0,312 | 0,448 | 0,572 | 58 | 0,377 | 0,610 | 0,672 | 93 | 0,513 | 0,707 | 0,792 |
| 24 | 0,313 | 0,451 | 0,575 | 59 | 0,382 | 0,610 | 0,673 | 94 | 0,530 | 0,714 | 0,793 |
| 25 | 0,313 | 0,458 | 0,576 | 60 | 0,382 | 0,612 | 0,676 | 95 | 0,586 | 0,715 | 0,800 |
| 26 | 0,313 | 0,460 | 0,588 | 61 | 0,383 | 0,617 | 0,677 | 96 | 0,600 | 0,716 | 0,831 |
| 27 | 0,314 | 0,472 | 0,598 | 62 | 0,385 | 0,622 | 0,677 | 97 | 0,606 | 0,732 | 0,839 |
| 28 | 0,317 | 0,489 | 0,602 | 63 | 0,396 | 0,624 | 0,685 | 98 | 0,610 | 0,733 | 0,851 |
| 29 | 0,321 | 0,489 | 0,610 | 64 | 0,399 | 0,625 | 0,698 | 99 | 0,617 | 0,735 | 0,856 |
| 30 | 0,326 | 0,502 | 0,621 | 65 | 0,412 | 0,627 | 0,698 | 100 | 0,630 | 0,756 | 0,865 |
| 31 | 0,328 | 0,503 | 0,621 | 66 | 0,415 | 0,627 | 0,704 | | | | |
| 32 | 0,329 | 0,528 | 0,627 | 67 | 0,416 | 0,634 | 0,707 | | | | |
| 33 | 0,330 | 0,532 | 0,628 | 68 | 0,418 | 0,635 | 0,708 | | | | |
| 34 | 0,331 | 0,540 | 0,639 | 69 | 0,422 | 0,640 | 0,715 | | | | |
| 35 | 0,332 | 0,541 | 0,640 | 70 | 0,427 | 0,642 | 0,723 | | | | |

Tablo 4, incelendiğinde, 10 puanlayıcılı örneklemlerde kontrol listesi kullanılarak puanlama yapıldığında (2 kategorili) güvenirlik kestirimleri 0,041 ile 0,630 arasında; analitik rubrik kullanılarak puanlama yapıldığında (3 kategorili) güvenirlik kestirimlerinin 0,214 ile 0,756 arasında; dereceleme ölçeği kullanılarak puanlama yapıldığında (5 kategorili) güvenirlik kestirimlerinin 0,278 ile 0,865 arasında değiştiği gözlenmektedir.

**II. Araştırma Sorusuna İlişkin Bulgular**

Araştırmanın ikinci sorusuna yanıt aramak amacıyla, G kuramı'na göre, aynı performans görevlerinin; kontrol listesi, dereceleme ölçeği, analitik rubrik kullanılarak ve 2, 3, 5, 10 puanlayıcıdan oluşan 100'er örneklemin puanlama güvenirliklerinin ortanca değerleri ve standart hataları elde edilmiştir. Bu değerler Tablo 5'te görülmektedir.

**Tablo 5.** Genellenebilirlik Kuram'ına İlişkin Güvenirliklerinin Ortancaları ve Standart Hataları

| Puanlayıcı sayısı | Ortanca/Standart Hata | 2 Kategorili (KL) | 3 Kategorili (AR) | 5 Kategorili (DPA) |
|---|---|---|---|---|
| 2 | Ortanca | 0,309 | 0,439 | 0,474 |
| | SH | 0,024 | 0,024 | 0,030 |
| 3 | Ortanca | 0,234 | 0,515 | 0,473 |
| | SH | 0,020 | 0,024 | 0,024 |
| 5 | Ortanca | 0,830 | 0,589 | 0,605 |
| | SH | 0,020 | 0,019 | 0,020 |
| 10 | Ortanca | 0,363 | 0,595 | 0,661 |
| | SH | 0,011 | 0,013 | 0,012 |

Tablo 5, 2 puanlayıcılı örneklemler açısından incelendiğinde, kontrol listesi kullanılarak puanlama yapıldığında (2 kategorili) G kuramı güvenirlik kestirimlerinin ortancalarının 0,309 olduğu, standart hataların 0,024 olduğu; analitik rubrik kullanılarak puanlama yapıldığında (3 kategorili) G kuramı güvenirlik kestirimlerinin ortancalarının 0,439 olduğu, standart hataların 0,024 olduğu; dereceleme ölçeği kullanılarak puanlama yapıldığında (5 kategorili) G kuramı güvenirlik kestirimlerinin ortancalarının 0,474 olduğu, standart hataların 0,030 olduğu gözlenmektedir.

Tablo 5, 3 puanlayıcılı örneklemler açısından incelendiğinde, kontrol listesi kullanılarak puanlama yapıldığında (2 kategorili) G kuramı güvenirlik kestirimlerinin ortancalarının 0,234 olduğu, standart hataların 0,020 olduğu; analitik rubrik kullanılarak puanlama yapıldığında (3 kategorili) G kuramı güvenirlik kestirimlerinin ortancalarının 0,515 olduğu, standart hataların 0,024 olduğu; dereceleme ölçeği kullanılarak puanlama yapıldığında (5 kategorili) G kuramı güvenirlik kestirimlerinin ortancalarının 0,473 olduğu, standart hataların 0,024 olduğu gözlenmektedir.

Tablo 5, 5 puanlayıcılı örneklemler açısından incelendiğinde, kontrol listesi kullanılarak puanlama yapıldığında (2 kategorili) G kuramı güvenirlik kestirimlerinin ortancalarının 0,830 olduğu, standart hataların 0,020 olduğu; analitik rubrik kullanılarak puanlama yapıldığında (3 kategorili) G kuramı güvenirlik kestirimlerinin ortancalarının 0,589 olduğu, standart hataların 0,019 olduğu; dereceleme ölçeği kullanılarak puanlama yapıldığında (5 kategorili) G kuramı güvenirlik kestirimlerinin ortancalarının 0,605 olduğu, standart hataların 0,020 olduğu gözlenmektedir.

Tablo 5, 10 puanlayıcılı örneklemler açısından incelendiğinde, kontrol listesi kullanılarak puanlama yapıldığında (2 kategorili) G kuramı güvenirlik kestirimlerinin ortancalarının 0,363 olduğu, standart hataların 0,011 olduğu; analitik rubrik kullanılarak puanlama yapıldığında (3 kategorili) G kuramı güvenirlik kestirimlerinin ortancalarının 0,595 olduğu, standart hataların 0,013 olduğu; dereceleme ölçeği kullanılarak puanlama yapıldığında (5 kategorili) G kuramı güvenirlik kestirimlerinin ortancalarının 0,661 olduğu, standart hataların 0,012 olduğu gözlenmektedir.

**TARTIŞMA ve SONUÇ**

Bu araştırmanın amacı, aynı performans görevinin farklı sayıda puanlayıcılar tarafından üç farklı teknikle puanlanmasından elde edilen puanların güvenirliklerinin G kuramı'na göre incelenmesidir.

Araştırmanın G kuramı'na dayalı olarak gerçekleştirilen puanlayıcılar arası güvenirlik çalışmaları birlikte değerlendirildiğinde, en yüksek uyumun beş puanlayıcı ve iki kategorili puanlama durumunda (G=0,960) elde edildiği belirlenmiştir.

Araştırmanın puanlayıcılar arası güvenirlik çalışmaları birlikte değerlendirildiğinde, puanlayıcılar arası güvenirliklerin ortancasının en yüksek olduğu durum G kuramı'na göre yine beş puanlayıcı ve iki kategorili puanlamanın olduğu durum (0,830) olarak karşımıza çıkmaktadır.

Elde edilen kestirimlerin iki kategorili puanlama dışında, puanlayıcı ve kategori sayısı arttıkça güvenirliğin arttığı gözlenmiştir. Alanyazında, yüzeylerden biri ya da her ikisindeki koşul sayısının arttırılması paydalardaki değerlerin azalmasına sebep olacağından güvenirlik katsayıları artmasına sebep olacağı ifade edilmektedir (Brennan, 2001; Güler, Uyanık ve Teker, 2012; Büyükkıdık, 2012). Dolayısıyla araştırmada elde edilen bu bulguların alanyazın ile paralellik gösterdiği görülmektedir.

Güvenirliklerin standart hataları incelendiğinde, en düşük standart hata değerinin (0,011) on puanlayıcı ve bu sefer iki kategorili puanlama sözkonusu olduğunda elde edildiği gözlenmiştir. G kuramı'nda puanlayıcı sayısı arttıkça standart hata değerlerinin düştüğü gözlenmiştir. Elde edilen bulgunun alanyazınla paralellik göstermektedir. Alanyazında standart hata ne kadar küçük olursa popülasyona ait tahminlerin o kadar isabetli olacağına ve örneklem sayısı arttıkça standart hata değerlerinin azalacağına yönelik açıklamalar mevcuttur (Özbek ve Keskin, 2007; Aşiret, 2014).

Elde edilen bulgular, puanlayıcı sayısı açısından incelendiğinde, iki puanlayıcı için G kuramı'nda kategori sayısı arttıkça güvenirliklerin ortancasının yükseldiği ve beş kategorili puanlamadan elde edilen güvenirliklerin ortancasının görece daha yüksek olduğu (0,474) belirlenmiştir. Üç puanlayıcı için G kuramı'nda yine üç kategorili puanlamadan elde edilen güvenirliklerin ortancasının görece daha yüksek olduğu (0,515) belirlenmiştir. Beş puanlayıcı için G kuramı'nda iki kategorili puanlamadan elde edilen güvenirliklerin ortancasının görece daha yüksek olduğu (0,830) ortaya konulmuştur. On puanlayıcı için G kuramı'nda kategori sayısı arttıkça güvenirliklerin ortancasının yükseldiği ve beş kategorili puanlamadan elde edilen güvenirliklerin ortancasının görece daha yüksek olduğu (0,515) gözlenmiştir.

Elde edilen bulgular, kategori sayısı (kontrol listesi, analitik rubrik, dereceleme ölçeği) açısından incelendiğinde, G kuramı'nda beş puanlayıcı olması durumunda elde edilen güvenirliklerin ortancasının görece daha yüksek olduğu (0,830) belirlenmiştir. Üç kategorili puanlamalar için G kuramı'nda puanlayıcı sayısı arttıkça güvenirliklerin ortancasının yükseldiği ve yine on puanlayıcı olması durumunda elde edilen güvenirliklerin ortancasının görece daha yüksek olduğu (0,595) belirlenmiştir. Üç kategorili puanlamalar için G kuramı'nda

puanlayıcı sayısı arttıkça güvenirliklerin ortancasının yükseldiği ve yine on puanlayıcı olması durumunda elde edilen güvenirliklerin ortancasının görece daha yüksek olduğu (0,661) belirlenmiştir. Puanlayıcı sayısı 5 ve kategori sayısı 2 olduğunda G kuramı'nda güvenirlik kestiriminin en yüksek değeri verdiği söylenebilir.

Sonuç olarak, bu çalışmada ortaya çıkan bulgular ışığında, G kuramı güvenirlik kestirimlerinin ortanca değerleri incelendiğinde 5 puanlayıcının kontrol listesi kullanarak yaptıkları puanlamaların güvenirliklerinin ortanca değeri hariç olmak üzere puanlayıcı sayısı ve aynı zamanda kullanılan ölçeğin kategori sayısı arttıkça ortanca değerlerinin de arttığı gözlenmiştir. G kuramı'ndan elde edilen standart hataların puanlayıcı sayısı arttıkça azaldığı gözlenmiştir. En düşük standart hata değerlerinin 10 puanlayıcı olması durumunda elde edildiği saptanmıştır. Puanlayıcı sayısı 5 ve kategori sayısı 2 olduğunda G kuramı'nda güvenirlik kestiriminin en yüksek değeri verdiği söylenebilir.

**ÖNERİLER**

Bu araştırmada elde edilen sonuçlara bağlı olarak araştırmacılar puanlayıcılar arası güvenirlik çalışmalarında G kuramına dayalı olarak iki kategorili ve beş puanlayıcının söz konusu olduğu çalışmalar planlayabilirler. Bu araştırma, 2, 3, 5 ve 10 puanlayıcılı örneklemlerle gerçekleştirilmiştir. Benzer çalışmalar yapacak olan araştırmacılar, yapılacak olan karar çalışmalarının ışığında farklı değerlendirici sayıları kullanarak araştırmalarını düzenleyebilirler. Araştırmada kullanılan farklı puanlama anahtarlarının mad¬de sayıları sabittir fakat farklı madde sayıları kullanarak da güvenirlik katsayıları incelenebilir. Bu araştırmada kullanılan üç farklı puanlama anahtarı dışında farklı sayıda ve türde ölçme araçları kullanılabilir. Bu araştırma kapsamında güvenirliğin değerlendirilmesinde, G kura¬mının çok değişkenli tümüyle çaprazlanmış g x pa x p deseni ile analizi yapılmıştır. Bu desen kullanılarak yapılan çalışmalarda iç geçerlik sorunu olabileceği düşüncesiyle benzer ça¬lışmaların yuvalanmış desen kullanılarak incelenmesi alana oldukça yararlı bilgiler sağlayacaktır.

**KAYNAKÇA**

Aiken, L. R. (2000). *Psychological Testing and Assessment (10th ed).* USA: Allyn and Bacon.

Anastasi, A. ve Urbina, S. (1997). *Psychological Testing (7th ed.)*. USA: Macmillan Pub. Co. Inc.

Arıcı, H,. (2005). *İstatistik: Yöntem ve Uygulamalar* (15. Baskı). Ankara: Meteksan A.Ş.

Aşiret, S,. (2014). Küçük Örneklemlerde Test Eşitleme Yöntemlerinin Çeşitli Faktörlere Göre İncelenmesi. Yayınlanmış Yüksek Lisans Tezi, Mersin Üniversitesi, Mersin.

Atılgan, H., (2005). G Kuramı ve Puanlayıcılar Arası Güvenirlik İçin Örnek Bir Uygulama. *Eğitim Bilimleri ve Uygulama*, 4 (7), 95-108.

Atılgan, H., (2004). G Kuramı *ve Çok Değişkenlik Kaynaklı Rasch Modelinin Karşılaştırılmasına İlişkin Bir Araştırma*. Yayımlanmış Doktora Tezi. Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü Eğitim Bilimleri Anabilim Dalı Eğitimde Ölçme ve Değerlendirme Bilim Dalı, Ankara.

Brennan, R., L. (2001). *Generalizability Theory*. USA: Springer-Verlag New York Inc.

Büyükkıdık. S,. (2012). *Problem Çözme Becerisinin Değerlendirilmesinde Puanlayıcılar Arası Güvenirliğin Klasik Test Kuramı ve G Kuramına Göre Karşılaştırılması.* Yayımlanmış Yüksek Lisans Tezi. Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü Eğitim Bilimleri Anabilim Dalı Eğitimde Ölçme ve Değerlendirme Bilim Dalı, Ankara.

Büyükkıdık. S,. Anıl, D,. (2015). Performansa Dayalı Durum Belirlemede Güvenirliğin Genellenebilirlik Kuramında Farklı Desenlerle İncelenmesi. *Eğitim ve Bilim*. 40 (177).

Covill. A. E. (2012). College Students' Use of a Writing Rubric: Effect on Quality of Writing, Self-Efficacy, and Writing Practices. The Journal of Writing Assessment. (5)1.

Çakıcı Eser. D,. Gelbal. S,. (2013). Genellenebilirlik Kuramı ve Lojistik Regresyona Dayalı Hesaplanan Puanlayıcılar Arası Tutarlığın Karşılaştırılması. *Kastamonu Eğitim Dergisi*. 21(2).

Deliceoğlu, G. (2009). *Futbol Yetilerine İlişkin Dereceleme Ölçeğinin Genellenebilirlik ve Klasik Test Kuramına Dayalı Güvenirliklerinin Karşılaştırılması.* Yayımlanmış Doktora Tezi. Ankara Üniversitesi Eğitim Bilimleri Enstitüsü Eğitimde Psikolojik Hizmetler Anabilim Dalı Ölçme ve Değerlendirme Bilim Dalı, Ankara.

Deliceoğlu, G., Çıkrıkçı Demirtaşlı. N,. (2012). Futbol Yetilerine İlişkin Dereceleme Ölçeğinin Genellenebilirlik ve Klasik Test Kuramına Dayalı Güvenirliklerinin Karşılaştırılması. *Hacettepe Spor Bilimleri Dergisi.* 23 (1),1–12.

Erkuş. A. (2006). Sınıf Öğretmenleri İçin Ölçme ve Değerlendirme: Kavramlar ve Uygulamalar. Ankara: Ekinoks.

Güler, N. (2011). Rasgele Veriler Üzerinde Genellenebilirlik Kuramı ve Klasik Test Kuramı'na Göre Güvenirliğin Karşılaştırılması. *Eğitim ve Bilim*, 36 (162).

Güler, N., Uyanık, G. K., Taşdelen Teker, G. (2012). *Genellenebilirlik Kuramı*. Ankara: Pegem Akademi.

Hobart, C., and Frankel, J. (1999) *A Practical Guide to Child Observation and Assessment*. Cheltenham: Stanley Thornes.

Keeves, J. P. (1988). *Educational Research, Methodology, and Measurement: an İnternational Handbook*. USA: Pergamon Press.

Kolen, M. J.,& Brennan R. L. (2004). Test Equating, Scaling, and Linking: Method and Practice (2nd ed.). New York, NY: Springer-Verlag.

Kothari. C. R. (2004). Research Methodology. New Delhi: New Age International (P) Ltd., Publishers.

Kutlu. Ö., Doğan. C. D., Karakaya. İ. (2008). Öğrenci Başarısının Belirlenmesi: Performansa ve Prtfolyoya Dayalı Durum Belirleme. Ankara: Pegem Akademi.

Moskal, Barbara M. & Jon A. Leydens (2000). Scoring Rubric Development: Validity and Reliability. *Practical Assessment, Research & Evaluation*. http://PAREonline.net/getvn.asp?v=7&n=10 web adresinden 24 Şubat 2012 tarihinde edinilmiştir.

Nalbantoğlu Yılmaz. F,. ve Gelbal. S,. (2011). İletişim Becerileri İstasyonu Örneğinde Genellenebilirlik Kuramı ile Farklı Desenlerin Karşılaştırılması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*. 41:509-518.

Nalbantoğlu Yılmaz. F,. ve Başusta. B,. (2015). Genellenebilirlik Kuramıyla Dikiş Atma ve Alma Becerileri İstasyonu Güvenirliğinin Değerlendirilmesi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*. 6(1).

Özbek, K. ve Keskin, S. (2007). Standart Sapma Mı Yoksa Standart Hata Mı? *Van tıp dergisi*. 14(2):64-67.

Popham, J. W. (1997). What's Wrong and What's Right With Rubric. *Educational Leadership.* 55, (2), 12.

Singer. N. B., LeMahieu. P. (2011). The Effect of Scoring Order on the Independence of Holistic and Analytic Scores. The Journal of Writing Assessment. (4)1.

Singh. Y. K. (2006). Fundamental of Research Methodology and Statistics. New Delhi: New Age International (P) Ltd., Publishers.

Tekindal. S, (Editör). (2008). *Eğitimde Ölçme ve Değerlendirme.* Ankara: Pegem Akademi.

Turgut, M. F., Baykul, Y. (2010). *Eğitimde Ölçme ve Değerlendirme (2. baskı).* Ankara: Pegem Akademi.