ISSN: 1309 – 6575
**Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi**
Journal of Measurement and Evaluation in Education and Psychology
2017; 8(1);63-78

# Puanlayıcılar Arası Güvenirlik Belirleme Tekniklerinin Karşılaştırılması*

# The Comparison of Interrater Reliability Estimating Techniques

Özge BIKMAZ BİLGEN **  Nuri DOĞAN ***

**Öz**

Bu araştırmada dereceli puanlama anahtarı türü ve puanlayıcı sayısı değişiminin, puanlayıcı güvenirliğini belirlemede kullanılan tekniklerden elde edilen sonuçlar üzerindeki etkisi incelenmiştir. Araştırmanın çalışma grubu, 50 öğrenci ve puanlama yapan 10 öğretmenden oluşmaktadır. Betimsel nitelik taşıyan araştırmada puanlayıcı güvenirliğini belirlemede Kappa istatistik tekniği, log linear analiz tekniği ve Krippendorff alfa tekniği kullanılmıştır. Puanlayıcı sayısı değişiminin puanlayıcı güvenirliğine etkisini incelemek adına belirtilen üç teknik kullanılarak iki, beş ve on puanlayıcı arasındaki uyum düzeyleri hesaplanmıştır. Araştırmada üç teknikten elde edilen analiz sonuçlarında, analitik puanlama anahtarı kullanımıyla elde edilen puanlarda, puanlayıcı sayısı artışının güvenirlik düzeyini düşürdüğü tespit edilmiştir. Üç teknikle yapılan analizlerde, en yüksek güvenirlik değerleri iki puanlayıcı kullanıldığında elde edilmiş, puanlayıcı sayısı artırıldıkça güvenirliğin düştüğü saptanmıştır. Analitik puanlama anahtarını oluşturan kategoriler incelendiğinde kategoriler arasında objektiflik düzeyine dayalı olarak, puanlayıcıların uyum düzeylerinde değişkenlik olduğu saptanmıştır. Araştırmanın sonucunda, kullanılan tekniklerden Kappa tekniği ve Krippendorff alfa tekniğinin paralel sonuçlar verdiği görülmüştür. Bununla birlikte Krippendorff alfa tekniğinin puanlayıcı sayısı değişiminden Kappa tekniğine göre daha az etkilendiği belirlenmiştir. Log-linear analiz tekniğinin ise değişkenler arasındaki etkileşimleri ve uyumsuzluk kaynağını gösteren daha kapsamlı ve geniş bilgi sağladığı tespit edilmiştir. Sonuç olarak, daha detaylı ölçme sonuçları elde edilmek istendiğinde alt kategorilerden oluşan analitik puanlama anahtarı kullanılarak toplanan puanların, kategorik veri analizi için uygun olan log-linear analiz tekniğinin; daha genel ölçme sonuçlarına ulaşılmak istendiğinde ise bütünsel puanlama anahtarı ile elde edilen puanların Krippendorff alfa tekniğinin kullanılmasının uygun olduğu düşünülmektedir.

*Anahtar Kelimeler:* Kappa istatistiği, log-linear analiz tekniği, Krippendorff alfa

**Abstract**

The aim of this study is to analyse the effects of the number of raters and the types of rubric on the results obtained by the techniques used to estimate the interrater reliability. The research group consists of 50 students and 10 teachers who rated. As a descriptive study, in this paper the Kappa statistical technique, the log linear analysis technique, and the Krippendorff alpha technique were used to determine the rater reliability. In order to investigate the effects of the number of raters on the interrater reliability, the level of agreement between 2, 5, and 10 raters was calculated by using those three techniques. The findings obtained from the three techniques demonstrated that the use of analytic rubric provided much more reliable ratings than holistic rubric. Moreover, it was also found based on the analysis results obtained through all three techniques that maximum reliability values were obtained by using two raters, reliability values decreased with the increase in the number of raters. On examining the categories constituting analytic rubric, it was found that there was variability in the levels of raters' agreement on the basis of objectivity. It was observed from the results that Kappa statistics and Krippendorff Alpha techniques yielded similar results. Moreover, Krippendorff alpha technique was found to be affected less by the number of raters. Log linear analysis technique, on the other hand, provided more comprehensive and extensive knowledge through showing the source of disagreement and interaction among the

*Bu makale, birinci yazar tarafından ikinci yazar danışmanlığında hazırlanan "Üst düzey zihinsel özelliklerin ölçülmesinde planlayıcılar arası güvenirlik belirleme tekniklerinin karşılaştırılması" başlıklı yüksek lisans tezinden üretilmiştir.

**Arş. Gör. Dr., Adnan Menderes Üniversitesi, Eğitim Fakültesi, Temel Eğitim Bölümü, Aydın-Türkiye, e-posta: ozgebikmaz@adu.edu.tr.

*** Prof. Dr., Hacettepe Üniversitesi, Eğitim Bilimleri Enstitüsü, Eğitimde Ölçme ve Değerlendirme Bilim Dalı, Ankara-Türkiye, e-posta:nurid@hacettepe.edu.tr



*Geliş Tarihi: 23.02.2017*
DOI: 10.21031/epod.294847
*Kabul Tarihi: 08.03.2017*

variants. As a result, it is thought that analyzing the scores obtained by using the analytic rubric which is composed of sub-categories using log-linear analysis technique would be more appropriate when the purpose is to obtain more detailed measurement results whereas analyzing the scores obtained through holistic rubric by using the Krippendorff technique would be more appropriate when the purpose is to obtain more general results.

*Keywords:* Kappa statistic, log linear analysis technique, Krippendorff alpha

## GİRİŞ

Ölçme ve değerlendirme eğitim sisteminin ayrılmaz ve önemli bir parçasıdır. Eğitim sisteminde ölçme ve değerlendirme uygulamalarından, eğitim sürecinin başında, süreç devam ederken ve sonunda yararlanılmaktadır. Hedeflenen istendik davranışların gerçekleşip gerçekleşmediğini belirleme, davranışta ne derecede bir değişim olduğunu saptama, süreçte kullanılan tekniklerin etkililik düzeylerini ve öğrenme sürecinde aksayan yönleri ortaya koyma amacıyla ölçme ve değerlendirme uygulamalarından yararlanılmaktadır. Ölçme ve değerlendirme sayesinde öğretim programının etkililiği değerlendirilebilir ve öğrencilerin öğrenme eksiklikleri ortaya konarak onların başarıları belirlenebilir (Atılgan, Kan ve Doğan, 2007).

Geleneksel ölçme yaklaşımları ile yapılan ölçmelerde, öğrencilerden, sınırlı bir zaman diliminde, kimseye danışmadan ya da belli kaynaklara başvurmadan verilen ölçme araçlarındaki soruları yanıtlaması beklenmektedir. Oysa bu tür bir ölçme sonucu gerçek yaşamda, öğrencilerin ölçme aracında verdiği yanıtlara uygun davrandığını kanıtlamadan uzaktır. Kutlu, Doğan ve Karakaya (2009) geleneksel değerlendirme yaklaşımlarının gerçek yaşam durumlarından uzak olmasının bir eleştiri noktası olduğunu, öğrencilerin çoktan seçmeli testlerle ya da boşluk doldurarak ne öğrendiklerinin değerlendirilmesinin çok zor olduğunu belirtmiştir. Geleneksel ölçme araçlarının belirtilen sınırlılıkları ile ölçülemeyen davranışlar performansa dayalı durum belirlemeyi gündeme getirmiştir.

Performansa dayalı durum belirleme Fitzpatrick ve Morrison (1971) tarafından "gerçek yaşam durumlarına benzer ortamlarda bireyin verdiği bir dizi yanıtın ölçülmesi" olarak tanımlanmaktadır. Bir dizi yanıt ile anlatılmak istenen, bireylerin ifade ettiği, yaptığı ya da ürettiğiyle ilgili davranışları içeren yanıtlardır. Bu bağlamda performanstan kasıt, bireyin ona sunulan öğrenme ortamında nasıl davrandığı ya da nasıl hareket ettiği ile ilgilenmektir. Eğitim alanında performans kavramıyla ilgili çok sayıda tanım yapılmıştır. Performansa dayalı durum belirleme süreci bağlamında performans, Kutlu ve diğerleri (2009) tarafından üst düzey zihinsel süreçleri gerektiren beceri ve yeteneklerle ilişkili olarak açıklanmıştır. Performans, beceri ve yetenekleri kapsayan karmaşık bir yapı olarak ele alınmaktadır. Performansa dayalı durum belirleme süreci bir dizi aşamayı gerektirmektedir. Bu aşamalar önceden belirlenmiş uygun bir görevin tanımlanması, yanıtlayıcıya verilmesi, yanıtlayıcı tarafından yerine getirilmesi ve sürecin gözlenerek puanlaması aşamalarını kapsamaktadır. Bu süreçte belirtilen puanlama aşamasında klasik ölçme araçlarından farklı olan ve performansın yapısına uygun ölçme araçları kullanılmaktadır.

Performansa dayalı durum belirleme sürecinde puanlama için kontrol listeleri ve puanlama ölçekleri kullanılmaktadır. Puanlama ölçeklerinden dereceli puanlama anahtarları bu süreçte sıklıkla kullanılan araçlarındandır (Airasian, 1994). Dereceli puanlama anahtarları, bir göreve ilişkin ölçüt listesini ve bu ölçütlere ilişkin niteliklerin derecesini içeren puanlama araçları olarak tanımlanmaktadır (Goodrich, 1997). Bu tür puanlama araçlarının, öğrencinin gelişmesi için dönüt sağlamak ya da öğrencinin belirli ölçütlere göre, ulaştığı düzeyi betimleyebilmek adına yararlı oldukları görülmektedir. Dereceli puanlama anahtarları alan yazında bütünsel (holistik) ve analitik olarak iki türde karşımıza çıkmaktadır (Mertler, 2001).

Analitik dereceli puanlama anahtarları, gözlemlere ait puanların, tanımlanmış kategorilerden (ölçütlerden) uygun düşen boyuta kaydedilmesini sağlayan ölçme araçlarıdır (Haladyna, 1997). Analitik puanlama anahtarı kullanımının ölçülecek performans çok boyutlu ve bileşenlerine ayrılabildiği durumlarda, performans görevine ilişkin öğretmene ve öğrenciye anlamlı geri bildirimler sağlaması istendiğinde, performansı belirlemek için yeterli süre olduğunda, üst eğitim kademelerinde kullanıldığı durumlarda yararlı oldukları belirtilmektedir (Mertler, 2001; Nitko, 2001). Bütünsel puanlama anahtarında ise analitik puanlama anahtarından farklı olarak öğrencinin gösterdiği

performans bütün olarak belirlenmekte ve öğrenciye tek bir puan verilmektedir (Kutlu ve diğerleri, 2009). Öğrencilerin performansı, parçalara ayrılmadan bütün olarak belirlenmek istendiği durumlarda kullanıldığı ifade edilmektedir (Moskal, 2000). Yani bu tür dereceli puanlama anahtarı öğrenci çalışmalarını bir bütün olarak ele alıp değerlendirmeyi amaçlamaktadır (Korkmaz, 2004). Bu nedenle bütünsel puanlama anahtarı performansın oluşumunda etkili olan süreçten çok ortaya çıkan ürüne, ürünün niteliğine ve sonucuna odaklanmaktadır (Atılgan, Kan ve Doğan, 2007).

### *Performansa Dayalı Durum Belirlemede Güvenirlik*

Öğrencinin sergilediği performansın olduğu gibi kabul edilmesi olanaksızdır. Diğer ölçme araçlarında olduğu gibi performansın belirlenmesinde kullanılan araçların da geçerli ve güvenilir olması gerekmektedir. Güvenirlik, ölçme sonuçlarının tesadüfi hatalardan arınıklığı olarak tanımlanmaktadır (Baykul, 2000). Tesadüfi hata ölçme işlemine hangi kaynaktan, ne derece karıştığı belli olmayan hata türüdür. Ölçmeyi yapanın dikkatsizliği, ölçme ortamı, ölçme aracı ve diğer faktörler bu tür hataya neden olabilir (Atılgan, Kan ve Doğan, 2007). Cohen, Swerdlik ve Phillips (1996) bu tür hatayı ölçme aracının hazırlanması, uygulanması, puanlanması ve yorumlanması aşamalarında ölçmeye karışan istenmedik durumlar olarak açıklamıştır. Ölçmeye karışan bu istenmedik durumları saptamak, hatasız sonuçlar elde etmek için araştırmacılar hata kaynaklarının ölçme sonuçları üzerindeki etkilerini hesaplamaya yönelik güvenirlik belirleme tekniklerini önermişlerdir. Güvenirlik belirleme teknikleri, test tekrar test, eşdeğer formlar olarak iki uygulamaya dayalı yöntemler ile Cronbach Alfa, KR-20-21, iki yarı yöntemi gibi içtutarlılık katsayısı olarak ifade edilen ve tek uygulamaya dayalı yöntemler olarak sınıflanmaktadır (Crocker ve Algina, 1986). Performansa dayalı durum belirlemede puanlayıcılar önemli bir hata kaynağı olarak sayılmaktadır. İstenmeyen değişkenlik kaynağının puanlayıcılar olduğu performansa dayalı durum belirleme sürecinde güvenirlik kestirimi için puanlayıcılar arası güvenirlik belirleme teknikleri önerilmiştir (Cohen, Swerdlik ve Phillips, 1996). Diğer bir deyişle bu tür ölçme araçlarının güvenirliği puanlayıcı kanısına dayalı olarak elde edilmektedir.

### *Puanlayıcılar Arası Güvenirlik*

Puanlayıcı güvenirliği puanlayıcı-içi ve puanlayıcılar-arası güvenirlik olarak iki türde incelenmektedir. Puanlayıcı-içi güvenirlik, aynı bireyin verdiği puanların birbiriyle tutarlılığı incelenerek hesaplanmaktadır. Çoğu araştırmada Cronbach Alfa katsayısı ile kestirilmektedir (Jonsson ve Svingby, 2007). Puanlayıcılar-arası güvenirlik ise birden fazla puanlayıcının verdiği puanlar arasındaki uyumun belirlenmesiyle hesaplanmaktadır. İki ya da daha fazla puanlayıcı (değerlendirici) arasındaki uyum veya tutarlılığın derecesi olarak tanımlanmaktır (Cohen ve diğerleri, 1996).

Puanlayıcılar arası güvenirlik aynı özelliği puanlayan birbirinden bağımsız iki ya da daha fazla puanlayıcı olduğunda herhangi bir durumda hesaplanabilmektedir (Viera ve Garret, 2005). Elde edilen güvenirlik değeri puanlayıcıların belli bir davranışın puanlanmasında ne derece fikir birliği içinde olduklarını yansıtmaktadır (Burry-Stock, Shaw, Laurie ve Chissom, 1996). Puanlayıcılar arası güvenirlik, puanlamanın bir puanlayıcıdan diğerine değişmemesi olarak tanımlanmaktadır (Kutlu ve diğerleri, 2009).

### *Puanlayıcılar Arası Güvenirlik Belirleme Teknikleri*

Alan yazında puanlayıcılar arası güvenirlik belirlemede uyum yüzdesi, puanlayıcılar-arası korelasyon katsayısı, puanlar arasındaki farka dayalı ANOVA gibi çok sayıda teknik kullanıldığı görülmektedir (Jonsson ve Svingby, 2007). Güvenirlik belirlemede amaç tesadüfi hata kaynaklarını belirlemektir. Puanlayıcı güvenirliğinde tesadüfi hata kaynağı olarak puanlayıcılar ele alınmaktadır. Uyum yüzdesi ya da korelasyona dayalı hesaplamalarda puanlayıcıların tesadüfe dayalı uyumlulukları hesaplanmadığı için eleştirildiği çalışmalara rastlanmaktadır. Puanlayıcılar-arası güvenirlik kestirimi

için çeşitli avantajlara sahip çok sayıda teknik önerilmiş, tekniklerden Kappa istatistiği, Krippendorff Alfa katsayısı ve log-linear analiz teknikleri bu çalışma kapsamında incelenmiştir.

*Kappa istatistiği (κ)*

Puanlayıcılar arası güvenirlik belirlemede sıklıkla kullanılan Kappa istatistiği, Cohen (1960) tarafından önerilmiştir. Sınıflama düzeyinde puanlama yapan iki puanlayıcı arasındaki uyumun derecesini belirlemek için geliştirilmiştir (Cohen, 1960). İki puanlayıcı ile sınırlı kalan κ istatistiği, Fleiss (1971) tarafından ikiden fazla puanlayıcı arasındaki uyumu belirlemede kullanılabilmesi için genelleştirilmiştir (Fleiss, 1971). Kappa istatistiği bazı temel varsayımlara dayanmaktadır (Crawforth, 2001). Bu varsayımlar Brennen ve Prediger (1981) tarafından puanlama sürecinde kategorilenen nesne ya da bireylerin bağımsız olduğu, puanlayıcıların puanlamalarının birbirinden bağımsız olduğu, puanlamada kullanılan kategorilerin birbirinden bağımsız olduğu şeklinde ifade edilmiştir. Kappa istatistiğinin bir avantajı kolay hesaplanması ve pratik yorumlanmasıdır. Diğer ve en önemli avantajı ise şansla beklenen uyumu düzeltmeyi temel almasıdır. Şansla meydana gelen uyum puanlardaki tamamen tesadüfe dayalı oluşan benzerliktir. κ, puanlayıcılar arası gözlenen uyumun içinden şansa/tesadüfe dayalı uyumun çıkarılmasına dayalı olarak Eşitlik 1'de verilen formülle hesaplanmaktadır. $\bar{P}$ gözlenen uyumluluk oranı, $\bar{P}_e$ tesadüfî/şansla uyumluluk oranı olmak üzere kappa istatistiği (κ ) formülüyle hesaplanmaktadır (Sim ve Wright, 2005).

$$\kappa = \frac{\bar{P}-\bar{P}_e}{1-\bar{P}_e} \qquad \textit{(Eşitlik 1)}$$

Kappa istatistiği -1 ile +1 arasında değer almaktadır (Fleiss, 1971). κ'nın pozitif değerleri puanlayıcılar arasındaki uyumun şansla beklenen uyumdan daha fazla olduğunu, κ'nın negatif değerleri puanlayıcılar arasındaki uyumun şansla beklenenden daha az olduğunu göstermektedir (Von Eye ve Mun, 2005). Bu anlamda negatif değerler şansla beklenenin altındaki uyum düzeyini gösterdiği için dikkate alınmamaktadır (Goodwin, 2001). κ istatistiğinin yorumlanmasında Tablo 1'de Landis ve Koch (1977) tarafından önerilen uyum düzeyleri kullanılmaktadır.

Tablo 1. Kappa İstatistiğinin Yorumlanmasına İlişkin Değer Aralıkları

| κ | Uyumun Gücü |
|---|---|
| < 0,00 | Zayıf |
| 0,00 – 0,20 | Önemsiz |
| 0,21 – 0,40 | Düşük |
| 0,41 – 0,60 | Orta |
| 0,61 – 0,80 | Önemli |
| 0,81 – 1,00 | Çok Yüksek |

*Krippendorff Alfa katsayısı (α)*

Krippendorff (1995) tarafından Krippendorff Alfa istatistiği adlı bir uyum ölçüsü önerilmiştir. Bu katsayı ilk olarak içerik analizinde kodlayıcılar arasındaki uyumun ölçüsünü belirlemeye yönelik olarak geliştirilmiştir. Bir uyum istatistiği olarak puanlayıcılar arasındaki uyumu belirlemede de kullanılmaktadır (Krippendorff, 1995, 2004, 2007). Krippendorff Alfa (α) istatistiği çok çeşitli veri tiplerine uygulanabilmektedir. Her değişken için herhangi bir sayıdaki değere uygulanabilir. İki veya daha fazla puanlayıcı içeren verilere uygulanabilir. Herhangi bir ölçek türü (sınıflama, sıralı, aralık, oran) ile ölçülmüş verilere uygulanabilir ve farklı büyüklükteki (küçük veya büyük) örneklemlerde kullanılabilir. Ayrıca puanlamada eksik veri olduğu durumlarda da uygulanabilir (Krippendorff, 1995). Alfa istatistiği için öncelikle gözlenen uyumsuzluk ($D_0$) ve beklenen uyumsuzluk ($D_e$) hesaplanmaktadır. Gözlenen uyumsuzluğun beklenen uyumsuzluğa bölünmesiyle elde edilen değer'in 1'den çıkartılması sonucunda α elde edilmektedir. Krippendorff Alfa katsayısının formülü Eşitlik 2'de verildiği şekildedir:

$$\alpha = 1 - \frac{D_0}{D_e} \qquad \textbf{(Eşitlik 2)}$$

Krippendorff alfa istatistiğinin yorumlanmasında α=1 olması puanlayıcılar arasındaki uyumun mükemmel olduğunu, α= 0 ise tam uyumsuzluğu simgelemektedir. Şansa bağlı olarak puanlayıcılar uyumlu oldukları zaman $D_o = D_e$ (Yani α=0) olur. Bu durum uyumun olmadığına işaret etmektedir. α'nın negatif çıkması şansla beklenin altında bir uyum olduğunu göstermektedir. Yüksek düzeyde güvenirliğin elde edilmesi amaçlandığından yorumlamada negatif değerler dikkate alınmamaktadır. α istatistiğinin yorumlanmasında Tablo 2'de verilen Krippendorff (1995) tarafından önerilen uyum düzeyleri kullanılmaktadır.

Tablo 2. Krippendorff Alfa Katsayısının Yorumlanmasına İlişkin Değer Aralıkları

| α | Uyumun gücü |
|---|---|
| < 0,67 | Zayıf |
| 0,67 - 0,80 | Orta |
| 0,80 ≤ | Yüksek |

*Log-linear Analiz Tekniği*

Çok yönlü çapraz tablolardaki kategorik veriler arasındaki ilişkiyi analiz etmek için ki-kare istatistiğinin uygulanabildiği ve yetersiz kaldığı durumlarda çok yönlü tabloların analizini modeller aracılığıyla yapabilen bir teknik olarak önerilmiştir (Agresti, 1996). Puanlayıcılar arası güvenirlik belirleme tekniği olarak kullanımı Tanner ve Young (1985)'in bu tekniği puanlayıcılar arasındaki uyumu modelleme denemeleriyle alan yazına girmiştir. Diğer tekniklerden farklı olarak sıralı ya da gruplanarak kategorik hale dönüştürülen eşit aralık ve oran ölçeğindeki verilerin iki yönlü, çok yönlü ve iç içe çapraz tablolarında birlikte değişimleri ve değişkenlerin alt kategorileri ile arasındaki etkileşimlerini analiz etmede kullanılmaktadır. Bu teknikte gözlenen frekansların modelini oluşturmak önemlidir. Seçilen modele göre her bir kategori için frekanslar hesaplanır ve bunlar gözlenen frekanslarla karşılaştırılır. Eğer uyum yeterli değilse model red edilir. Diğer bir deyişle log-linear analizde k tane değişken analize alınıyorsa "k veya daha fazla dereceli etkiler 0'a eşittir" hipotezine yanıt aranır. Elde edilen değer istatistiksel açıdan önemliyse $H_0$ red edilir.

Log-linear analiz sonuçlarının yorumlanmasında modelin uygun olup olmadığına Eşittlik 3 ve 4'te verilen Pearson ki-kare ($\chi^2$) ve olasılık oranı $L^2$ ile karar verilmektedir.

$$\chi^2 = \sum \frac{(O_i - E_i)}{E_i} \qquad \textbf{(Eşitlik 3)}$$

ve

$$L^2 = 2 \sum O_i \ln\left(\frac{O_i}{E_i}\right) \qquad \textbf{(Eşitlik 4)}$$

Yukarıdaki formüllerde, $O_i$, gözlenen hücre frekansı, $E_i$, beklenen hücre frekansıdır. Eğer sonuç istatistiksel açıdan önemliyse, model uyumu zayıftır, $H_0$ hipotezi, dolayısıyla da model red edilir (Agresti ve Yang, 1987).

Yukarıda açıklanan tekniklerin avantajları ve sınırlılıkları mevcuttur. Bu çalışmada ilgili teknikler üzerinde dereceli puanlama anahtarı türü ve puanlayıcı sayısı değişiminin etkisinin incelenmesi için iki alt problem oluşturulmuştur.

1. İki, beş ve on puanlayıcı ile analitik dereceli puanlama anahtarı aracılığıyla puanlama yapıldığında, Kappa istatistiği, Krippendorff Alfa katsayısı ve Log linear analiz tekniği ile elde edilen güvenirlik sonuçları nasıldır?

2. İki, beş ve on puanlayıcı ile bütünsel puanlama anahtarı aracılığıyla puanlama yapıldığında, Kappa istatistiği, Krippendorff Alfa katsayısı ve Kendall uyum istatistiği ile elde edilen güvenirlik sonuçları nasıldır?

### *Araştırmanın Amacı*

Araştırmanın amacı, aynı amaca yönelik hazırlanan ve aynı bireylere uygulanan performansın analitik ve bütünsel puanlama anahtarı kullanılarak puanlanmasıyla Kappa, Krippendorf ve Log-Linear analiz teknikleriyle hesaplanan güvenirlik değerlerini incelemek ve karşılaştırmaktır. Aynı zamanda bu araştırmada, puanlayıcı sayısı değişiminin, tekniklere ve ölçme aracına bağlı olarak puanlayıcı güvenirliğinde bir değişime neden olup olmadığının belirlenmesi amaçlanmaktadır.

## YÖNTEM

### *Araştırmanın Türü*

Bu araştırma, Kappa istatistiği, Krippendorff Alfa katsayısı ve log-linear analiz tekniğinin uygulanmasına, bu tekniklerin benzerlik ve farklılıklarının belirlenmesine, sınırlılıklarının incelenmesine, tekniklerden hangisinin daha fazla bilgi sağladığının saptanmasına dayanmaktadır. Bu yönüyle durum saptamaya yönelik olduğu için betimsel bir araştırma niteliği taşımaktadır.

### *Çalışma Grubu*

Araştırmanın çalışma grubu, performans görevini yerine getiren 50 öğrenciden ve bu görevleri biri analitik diğeri bütünsel iki ayrı dereceli puanlama anahtarı kullanarak puanlayan 10 öğretmenden oluşmaktadır. Çalışma grubunda yer alan öğrenciler, araştırmacılardan ilkinin sınıf öğretmeni olarak görev yaptığı İstanbul ili Beyoğlu ilçesine bağlı bir devlet okulunun beşinci sınıfında öğrenim gören 50 öğrenciden oluşmaktadır. Öğrencilerin belirlenmesinde, öğrencilere kolay ulaşabilme, uygulama döneminde öğrencileri izleyebilme gibi kolaylıklar etkili olmuştur. Çalışma grubunda yer alan öğretmenler, Milli Eğitim Bakanlığı yapısındaki İstanbul ilinde sınıf öğretmeni olarak görev yapan öğretmenlerden oluşmaktadır. Öğretmenlerin seçiminde gönüllük esası benimsenmiştir.

### *Veri Toplama Araçları*

*Performans görevi*

İlköğretim fen ve teknoloji dersi öğretim programında yer alan "Vücudumuzun bilmecesini çözelim" ünitesiyle ilişkili Kutlu, Doğan ve Karakaya (2009) tarafından hazırlanan performans görevidir.

*Analitik dereceli puanlama anahtarı*

Kutlu ve diğerleri (2009) tarafından geliştirilen analitik puanlama anahtarından yararlanılarak oluşturulmuştur. Analitik puanlama anahtarı altı alt kategoriden oluşmaktadır. Bu kategoriler: içerik, araştırma süreci, materyal kullanımı, grafik oluşturma, tablo oluşturma ve zaman kullanımıdır. Beş ölçme ve değerlendirme uzmanı kanısı, çalışma koşulları düşünülerek zaman kullanımı alt kategorisi analize dahil edilmeden son şekliyle analitik puanlama anahtarı beş alt boyutlu olarak oluşturulmuştur. Her bir performans 1-3 arasında puanlanmaktadır. Alınabilecek en yüksek puan 15; en düşük puan 5'tir.

*Bütünsel dereceli puanlama anahtarı*

Kutlu ve diğerleri (2009) tarafından geliştirilen analitik puanlama anahtarından yararlanılarak oluşturulmuş, beş ölçme ve değerlendirme uzmanı kanılarına dayalı olarak son şekli verilmiştir. Bu araçtan alınabilecek en yüksek puan 4 iken en düşük puan 1'dir.

### *İşlem*

Çalışma kapsamında işlemler, performans görevinin uygulanması ve bu görevin puanlanması olarak iki aşamada gerçekleştirilmiştir.

*Performans görevinin uygulanması*

Öğrencilere performans görevi ve performanslarının değerlendirileceği ölçütler performans göreviyle birlikte yazılı olarak verilmiş, varsa anlamadıkları noktalar öğrencilere açıklanmıştır. Performans görevini tamamlamaları için öğrencilere 10 gün süre verilmiştir. 10 gün boyunca araştırmacının belirlediği zamanda öğrencilerle görüşülmüş, varsa soruları yanıtlanmıştır. Öğrencilere verilen süre sonlandığında performans görevlerine ilişkin raporlar toplanmış, görevlere araştırmacı tarafından puan verilmiştir.

*Performans görevinin puanlanması*

Puanlamayı yapan öğretmenlere performans görevi, dereceli puanlama anahtarları ve puanlama süreciyle ilgili genel bilgi verilmiştir. Puanlamada hatırlama etkisini ortadan kaldırmak adına öğrencilere 1'den 50'ye kadar numaralar verilerek kod numaraları oluşturulmuştur. Her öğretmen ayrı bir zaman diliminde puanlama yapmıştır. Puanlamada yine hatırlama etkisini ortadan kaldırmak adına her bir öğretmenin bütünsel dereceli puanlama anahtarı kullandığı ikinci puanlama ile ilk puanlama arasında 2 hafta süre bırakılmıştır. Puanlayıcıların aynı oturumda tüm öğrencilere puan verme işlemini tamamlaması sağlanmıştır.

### *Verilerin Analizi*

Verilerin çözümlenmesinde analitik puanlama anahtarı için puanlayıcılar arası güvenirlik belirleme tekniklerinden Kappa istatistiği, Krippendorff alfa katsayısı, log linear analiz tekniği kullanılmıştır. Log-linear analiz sadece kategorik verilere uygulanabildiği için bütünsel dereceli puanlama anahtarı ile elde edilen sonuçların analizinde kullanılamamıştır. Bütünsel dereceli puanlama anahtarı ile verilerin analizi için log-linear analiz yerine Kendall'ın uyum istatistiği kullanılmıştır. Bu haliyle bütünsel puanlama anahtarından elde edilen veriler için Kappa istatistiği, Krippendorff alfa katsayısı ve Kendall'ın uyum katsayısından yararlanılmıştır. Kappa istatistiğinin hesaplanmasında "SPSS syntax (mkappasc.sps)" dosyasından, Krippendorff için "SPSS syntax (kalpha.sps)" dosyasından, log linear analiz tekniği ve Kendall katsayıları için ise doğrudan SPSS paket programından yararlanılmıştır.

## BULGULAR

Araştırmadan elde edilen bulgulara dayalı olarak ulaşılan sonuçlar araştırma soruları doğrultusunda değerlendirilmiştir. Analitik dereceli puanlama anahtarı kullanılarak yapılan puanlamada Kappa istatistiğine ilişkin bulgular Tablo 3'te özetlenmiştir.

Tablo 3. Analitik Puanlama Anahtarı ile Yapılan Puanlamaların Kappa İstatistiğiyle Hesaplanan Güvenirlik Değerleri

| Puanlama Yapan Puanlayıcı Sayısı | Puanlama Anahtarının Kategorileri | Kappa İstatistiği Değeri (κ) |
|---|---|---|
| 2 | İçerik | 0,40* |
| | Araştırma Süreci | 0,59* |
| | Materyal Kullanımı | 0,65* |
| | Grafik Oluşturma | 0,81* |
| | Tablo Oluşturma | 0,92* |
| 5 | İçerik | 0,34* |
| | Araştırma Süreci | 0,57* |
| | Materyal Kullanımı | 0,64* |
| | Grafik Oluşturma | 0,78* |
| | Tablo Oluşturma | 0,84* |
| 10 | İçerik | 0,32* |
| | Araştırma Süreci | 0,52* |
| | Materyal Kullanımı | 0,64* |
| | Grafik Oluşturma | 0,75* |
| | Tablo Oluşturma | 0,81* |

**p<0,001*

Tablo 3'e göre iki, beş ve on puanlayıcının verdikleri puanlar arasındaki uyumu elde etmek amacıyla analitik puanlama anahtarının kategori değişkeninin her biri için hesaplanan Kappa değerleri istatistiksel açıdan anlamlı bulunmuştur ($p<0,001$). Burada κ istatistiklerinin pozitif olması, puanlayıcılar arasında tesadüfen çıkabilecek olası uyumdan daha yüksek düzeyde uyum olduğuna işaret etmektedir (Landis ve Koch, 1977). İki puanlayıcının olduğu koşul için Kappa değerleri 0,40 ile 0,92 arasındadır. Bu koşulda en düşük uyum, içerik kategorisinde elde edilmişken ($\kappa = 0,40$); en yüksek uyum, tablo oluşturma kategorisi için kestirilmiştir ($\kappa = 0,92$). Diğer üç kategori değerlendirildiğinde, araştırma süreci kategorisinde puanlayıcılar arasında orta düzeyde, materyal kullanımında önemli düzeyde, grafik oluşturmada ise çok yüksek düzeyde uyum olduğu görülmektedir.

Tablo 3'teki beş puanlayıcının olduğu koşul için κ istatistiğine ait değerlerin, 0,34 ile 0,84 arasında; on puanlayıcı olduğu koşulda ise 0,32 ile 0,81 arasında değişkenlik gösterdiği görülmektedir. İki, beş ve on puanlayıcının olduğu bulgular karşılaştırıldığında, Kappa istatistiği değerlerinin anlamlılık düzeyi ve değerlerinin kategori bazında κ değerlerinin büyükten küçüğe doğru sıralanışı değişmemiştir. Ancak Kappa değerleri puanlayıcı sayısı arttıkça görece azalmıştır. Tablo 3'te verildiği gibi analitik puanlama anahtarının kategorilerinin κ değerleri sırayla içerik, araştırma süreci, materyal kullanımı, grafik oluşturma, tablo oluşturma olarak artan değerler almıştır. "içerik" kategorisinden "tablo oluşturma" kategorisine doğru gidildikçe elde edilen Kappa istatistiklerinin değeri artmıştır. "Tablo oluşturma" kategorisinde öğrenciden oluşturacağı tabloda satır, sütun isimlerini vermesi ve gözeneklerde bilgi vermesi beklenmektedir. "İçerik" kategorisinde ise öğrenciden konuyla ilgili tanıtıcı, açıklayıcı ve kaynaklara dayalı bilgi vermesi istenmektedir. Bu kategoride öğrenciden beklenen performansın çerçevesi puanlayıcının konuyla ilgili bilgi düzeyine, konuyu algılayışına ve yorumlayışına göre değişebilir. Nitekim bazı puanlayıcılar öğrencilerden konuyla ilgili performansın ayrıntılı bir sunumunu, bazıları ise kısa ve net sunumunu beklemektedir. Puanlayıcıların nitelikli buldukları sunumlar kişiden kişiye değişmektedir. Yani kategoride yer alan ölçüt bir yönüyle puanlayıcıdan puanlayıcıya farklılık gösterebilmektedir. Hesaplanan Kappa istatistiği değerlerinin Tablo 3'teki sırasıyla "içerik" kategorisinden, "tablo oluşturma" kategorisine gidildikçe artması, "tablo oluşturma" kategorisine doğru kategorilerin daha objektif olmasına dayandırılabilir.

Araştırmanın ilk alt problemi kapsamında analitik dereceli puanlama anahtarı ile puanlama yapıldığında Krippendorff alfa istatistiği ile iki, beş ve on puanlayıcı olduğunda elde edilen puanlamaların güvenirliğine ait bulgular Tablo 4'te özetlenmiştir.

Tablo 4. Analitik Puanlama Anahtarı ile Yapılan Puanlamaların Krippendorff Alfa Katsayısıyla Hesaplanan Güvenirlik Değerleri

| Puanlama Yapan Puanlayıcı Sayısı | Puanlama Anahtarının Kategorileri | Alfa Katsayısı Değeri (α) |
|---|---|---|
| 2 | İçerik | 0,59 |
| | Araştırma Süreci | 0,74 |
| | Materyal Kullanımı | 0,79 |
| | Grafik Oluşturma | 0,87 |
| | Tablo Oluşturma | 0,93 |
| 5 | İçerik | 0,52 |
| | Araştırma Süreci | 0,75 |
| | Materyal Kullanımı | 0,78 |
| | Grafik Oluşturma | 0,86 |
| | Tablo Oluşturma | 0,89 |
| 10 | İçerik | 0,51 |
| | Araştırma Süreci | 0,71 |
| | Materyal Kullanımı | 0,78 |
| | Grafik Oluşturma | 0,86 |
| | Tablo Oluşturma | 0,87 |

Analitik puanlama anahtarı kullanılarak yapılan puanlamada, puanlama anahtarının kategorileri için hesaplanan Krippendorff Alfa katsayısına ilişkin bulgular Tablo 4'te verilmiştir. Tablo 4 incelendiğinde, iki puanlayıcının puanları arasındaki uyumu elde etmek için kestirilen Krippendorff alfa katsayısı değerleri 0,59 ile 0,93 arasındadır. En düşük uyum "içerik" kategorisinde çıkmışken (α= 0,59); en yüksek uyum "tablo oluşturma" kategorisi için elde edilmiştir (α= 0,93). "Araştırma süreci" ve "materyal kullanımı" kategorileri için orta düzeyde uyum; "grafik oluşturma" ve "tablo oluşturma" kategorileri için yüksek düzeyde uyum olduğu tespit edilirken, içerik kategorisinde zayıf uyum olduğu görülmektedir. Buna göre "içerik" dışındaki kategorilerden elde edilen puanların güvenilir olduğu söylenebilir.

Beş puanlayıcının puanları arasındaki uyumu için hesaplanan Krippendorff alfa değerleri 0,52 ile 0,89 arasındadır. En düşük uyum "içerik" kategorisinde (α = 0,52); en yüksek uyum "tablo oluşturma" kategorisi için elde edilmiştir (α = 0,89). On puanlayıcının olduğu koşullarda alfa değerleri 0,51 ile 0,87 arasında değişmiştir. İki, beş ve on puanlayıcının olduğu bulgular karşılaştırıldığında, Krippendorff alfa katsayısı değerlerinin kategori bazında büyükten küçüğe doğru sıralanışı (tablo oluşturma, grafik oluşturma, materyal kullanımı, araştırma süreci, içerik) şeklinde değişmemiştir. Ancak alfa değerleri puanlayıcı sayısı ikiden beşe, beşten ona çıkarıldığında görece azalmıştır.

Analitik dereceli puanlama anahtarı kullanılarak yapılan puanlamada log-linear analiz bulgularına ait değerler Tablo 5'te verilmiştir.

Tablo 5. Analitik Puanlama Anahtarı ile Yapılan Puanlamaların Log-linear Analiz Tekniğiyle Hesaplanan Güvenirlik Değerleri

| Puanlama Yapan Puanlayıcı Sayısı | Puanlama Anahtarının Kategorileri | *LL* $\chi^2$ | Etki Düzeyi |
|---|---|---|---|
| 2 | İçerik | 34,072* | 1. Düzey |
| | Araştırma Süreci<br>Materyal Kullanımı<br>Grafik Oluşturma | 56,233* | 2. Düzey |
| | Tablo Oluşturma | 7,759 | 3. Düzey |
| 5 | İçerik | 55,931* | 1. Düzey |
| | Araştırma Süreci<br>Materyal Kullanımı<br>Grafik Oluşturma | 144,333* | 2. Düzey |
| | Tablo Oluşturma | 36,636 | 3. Düzey |
| 10 | İçerik | 34,616* | 1. Düzey |
| | Araştırma Süreci<br>Materyal Kullanımı<br>Grafik Oluşturma | 314,451* | 2. Düzey |
| | Tablo Oluşturma | 79,774 | 3. Düzey |

*$p<0,05$

Log-linear analiz tekniğinde puanlayıcılar, kategoriler, alt kategoriler şeklinde üç değişken tanımlanmıştır. Öncelikle ilgili değişkenlerin tek başına etkilerinin yanında ikili ve üçlü etkilerinin birlikte bulunduğu logaritmik modelle ifade edilip edilemeyeceği incelenmiştir.

Kategoriler değişkeni analitik puanlama anahtarındaki "içerik, araştırma süreci, materyal kullanımı, grafik oluşturma, tablo oluşturma" olarak beş tanedir. Kategoriler değişkenin her birine 1 ile 3 arasında değişen puan verilerek alt kategoriler değişkeni tanımlanmıştır. Diğer bir ifadeyle alt kategoriler değişkeni kategoriler değişkenine verilen puanlara dayanarak oluşturulmuştur. Her kategori değişkeninin altında üç tane altkategori değişkeni mevcuttur. İki, beş ve on puanlayıcı olduğu durumlarda bu değişkenlerin tek başına, ikili ve üçlü etkilerinin birlikte bulunduğu koşulların logaritmik modelle ifade edilip edilemeyeceği sınanmış, tekli ve ikili etkilerin istatistiksel açıdan anlamlı ($p<0,05$); üçlü etkilerin ise anlamsız olduğu ($p>0,05$) görülmüştür. Anlamlı olan tekli ve ikili etkiler için gerçekleştirilen log-linear analiz bulguları Tablo 6'da verilmiştir.

Tablo 6. Log-linear Analiz Tekniğinden Elde Edilen Bulgulara İlişkin Değerler

| Puanlama Yapan Puanlayıcı Sayısı | Değişkenlerin Tekli ve İkili Etkileri | Serbestlik Derecesi | *Kısmi* $\chi^2$ |
|---|---|---|---|
| 2 | Puanlayıcı*kategori | 4 | 0,026 |
| | Puanlayıcı*altkategori | 2 | 0,486 |
| | Altkategori*kategori | 8 | 55,791* |
| | Puanlayıcı | 1 | 0,000 |
| | Kategori | 4 | 0,000 |
| | Altkategori | 2 | 34,072* |
| 5 | Puanlayıcı*kategori | 16 | 0,680 |
| | Puanlayıcı*altkategori | 8 | 11,693 |
| | Altkategori*kategori | 8 | 133,320* |
| | Puanlayıcı | 4 | 0,000 |
| | Kategori | 4 | 0,000 |
| | Altkategori | 2 | 55,931* |
| 10 | Puanlayıcı*kategori | 36 | 2,120 |
| | Puanlayıcı*altkategori | 18 | 50,054 |
| | Altkategori*kategori | 8 | 266,517* |
| | Puanlayıcı | 9 | 0,000 |
| | Kategori | 4 | 0,000 |
| | Altkategori | 2 | 34,616* |

*$p<0,05$

Tablo 6'da yer alan bulgular incelendiğinde, iki puanlayıcının olduğu koşulda, değişkenlerin tek başlarına etkilerinin anlamsız olduğu yönünde kurulan hipotezlerden puanlayıcı ve kategori değişkenleri için olanlar kabul edilmiştir. Diğer bir ifade ile bu değişkenlerde puanlayıcılar arasındaki farklılık manidar değildir ($p>0,05$). İçerik, araştırma süreci, materyal kullanımı, grafik oluşturma, tablo oluşturma şeklinde beş kategoriden oluşan kategoriler bazında puanlamalarda manidar farklılık yoktur. "Altkategori" değişkenin etkisinin anlamsız olduğu yönündeki hipotez testi ise red edilmiştir. Yani alt kategoriler bazındaki puanlamalar arasında anlamlı fark vardır ($p<0,05$). İki puanlayıcının olduğu durumda tekli etkilere göre bulgular puanlamalardaki farklılığın "altkategori" değişkeninden kaynaklanabileceğini göstermektedir ($\chi^2 = 34,07$; sd= 2; $p<0,05$).

Puanlayıcı sayısının iki olduğu koşulda, puanlamada değişkenlerin ikili etkileşimlerinin farklılıklarının anlamsız olduğu yönündeki hipotezlerden puanlayıcı*altkategori ve puanlayıcı*kategori etkileşimleri için kurulanlar kabul edilmiştir. Buna göre bu etkileşimlerin anlamlı olmadığı görülmektedir ($p>0,05$). Ancak altkategori*kategori etkileşiminin anlamsız olduğu yönünde kurulan hipotez red edilmiştir, yani bu etkileşimin puanlamada oluşturduğu farklılık anlamlı bulunmuştur ($p<0,05$). "Altkategori" değişkeninin hem tek başına hem de kategori değişkeniyle etkileşimi sonucunda farklılık oluşturması puanlardaki asıl değişkenlik kaynağının "altkategori" değişkeni olduğu şeklinde yorumlanabilir ($\chi^2 = 55,79$; sd=8; $p<0,05$).

Tablo 6'da beş puanlayıcı için log linear analiz bulguları incelendiğinde, puanlayıcı ve kategori değişkenlerinin tek başına etkilerinin anlamlı olmadığı; "altkategori" etkisinin ise anlamlı olduğu tespit edilmiştir ($\chi^2 = 55,931$; sd= 2; $p<0,05$). Değişkenlerin ikili etkileşimleri incelendiğinde ise puanlayıcı*altkategori ve puanlayıcı*kategori etkilerinin anlamsız; kategori*altkategori etkisinin ise anlamlı olduğu gözlenmiştir ($\chi^2 = 133,32$; sd=8; $p<0,05$). "Altkategori" değişkeninin hem tek başına hem de kategori değişkeniyle etkileşimi sonucunda farklılık oluşturması puanlardaki asıl değişkenlik kaynağının "altkategori" değişkeni olduğu şeklinde yorumlanabilir.

Tablo 6'da on puanlayıcı için log linear analiz bulguları incelendiğinde, puanlayıcı ve kategori değişkenlerinin tek başına etkilerinin anlamlı olmadığı; "altkategori" etkisinin ise anlamlı olduğu tespit edilmiştir ($\chi^2 = 34,616$; sd= 2; $p<0,05$). Değişkenlerin ikili etkileşimleri incelendiğinde ise puanlayıcı*altkategori ve puanlayıcı*kategori etkilerinin anlamsız; altkategori*kategori etkisinin ise anlamlı olduğu gözlenmiştir ($\chi^2 = 266,517$; sd=8; $p<0,05$). "Altkategori" değişkeninin hem tek başına hem de kategori değişkeniyle etkileşimi sonucunda farklılık oluşturması, puanlardaki asıl değişkenlik kaynağının "altkategori" değişkeni olduğu şeklinde yorumlanabilir.

Bütünsel dereceli puanlama anahtarı kullanılarak yapılan puanlamada Kappa istatistiğine ilişkin bulgular Tablo 7'de özetlenmiştir.

Tablo 7. Bütünsel Puanlama Anahtarı ile Yapılan Puanlamaların Kappa İstatistiğiyle Hesaplanan Güvenirlik Değerleri

| Puanlama Yapan Puanlayıcı Sayısı | Kappa İstatistiği Değeri (κ) |
|---|---|
| 2 | 0,38* |
| 5 | 0,24* |
| 10 | 0,27* |

*$p<0,001$

Tablo 7'ye göre iki puanlayıcının bütünsel dereceli puanlama anahtarı ile yaptıkları puanlamaların uyumu için kullanılan Kappa istatistiğinin değeri 0,38; beş puanlayıcı için 0,24; on puanlayıcı için ise 0,27 olarak hesaplanmıştır. Buna göre bütünsel puanlama anahtarı ile yapılan puanlamalarda Kappa değeri ile puanlayıcıların birbiriyle düşük düzeyde uyum gösterdikleri görülmektedir.

Bütünsel dereceli puanlama anahtarı kullanılarak yapılan puanlamalar arasındaki uyum için hesaplanan Krippendorff alfa istatistiğine ilişkin bulgular Tablo 8'de özetlenmiştir.

Tablo 8. Bütünsel Puanlama Anahtarı ile Yapılan Puanlamaların Krippendorff İstatistiğiyle Hesaplanan Güvenirlik Değerleri

| Puanlama Yapan Puanlayıcı Sayısı | Alfa Katsayısı Değeri (α) |
|---|---|
| 2 | 0,67 |
| 5 | 0,60 |
| 10 | 0,58 |

Tablo 8'e göre iki puanlayıcının bütünsel puanlama anahtarı ile verdiği puanlar arasındaki Krippendorf alfa istatistiği ile hesaplanan uyum değerleri 0,67; beş puanlayıcı ile hesaplanan uyum 0,60; on puanlayıcı ile hesaplanan uyum değerleri ise 0,58 olarak belirlenmiştir. Bu değerler, puanlayıcı arasındaki uyumun zayıf düzeyde olduğunu göstermektedir.

Tablo 9. Bütünsel Puanlama Anahtarı ile Yapılan Puanlamaların Kendall Uyum Katsayısıyla Hesaplanan Güvenirlik Değerleri

| Puanlama Yapan Puanlayıcı Sayısı | Kendall Katsayısı Değeri (w) |
|---|---|
| 2 | 0,61* |
| 5 | 0,31* |
| 10 | 0,18* |

*$p<0,001$

Tablo 9'a göre iki puanlayıcının puanları arasındaki uyumu elde etmek için hesaplanan Kendall uyum istatistiği 0,61 olarak hesaplanmış ve bu değer istatistiksel açıdan anlamlı bulunmuştur ($p<0,001$). Sıra farklarını dikkate alarak uyumun hesaplandığı bir teknik olan Kendall'ın uyum istatistiği iki puanlayıcı için orta düzeyde çıkmıştır. Bu bulgu puanlayıcıların bireyleri sıralamada farklılık gösterdiği şeklinde yorumlanabilir. Beş puanlayıcının puanları arasındaki uyumu elde etmek için kullanılan Kendall'ın uyum istatistiği 0,31 olarak hesaplanmış ve bu değer anlamlı çıkmıştır ($p<0,001$). Bu değerin, Von Eye ve Mun (2005)'un ölçütlerine göre düşük düzeyde olduğu söylenebilir. On puanlayıcının puanları arasındaki uyumu elde etmek için hesaplanan Kendall'ın uyum istatistiği istatistiksel açıdan anlamlı bulunmuştur ($p<0,001$). Hesaplanan değer 0,18 olup, bu değer uyumun zayıf düzeyde olduğunu göstermektedir. Sıra farklarını dikkate alarak uyumun hesaplandığı teknik olan Kendall istatistiğinin çok düşük olması puanlayıcıların bireyleri sıralamada farklılık gösterdiği şeklinde yorumlanabilir.

## SONUÇLAR ve TARTIŞMA

Aynı amaca yönelik analitik puanlama anahtarı ile yapılan puanlama, bütünsel puanlama anahtarı kullanılarak yapılan puanlamaya göre göreceli olarak daha objektif sonuçlar vermiştir. Puanlayıcılar arasında daha standart ve daha nesnel sonuçlar veren analitik dereceli puanlama anahtarının daha tutarlı puanlama sağladığı, dolayısıyla daha güvenilir olduğu sonucuna varılmıştır. Bu sonuç, Kutlu ve diğerleri (2009)'nin bütünsel dereceli puanlama anahtarından elde edilen sonuçların analitik puanlama anahtarından elde edilen sonuçlara göre güvenirlik düzeyinin düşük olduğu yönündeki açıklamalarıyla örtüşmektedir. Aynı zamanda bu çalışmanın sonuçları, Jonsson ve Svingby (2007)'ın çalışmasının analitik puanlama anahtarının güvenirliği artırdığı yönündeki sonuçlarını desteklemektedir.

Araştırma kapsamında, alt kategorilere sahip analitik puanlama anahtarı kullanılarak elde edilen sonuçlar incelenmiştir. Puanlama anahtarını oluşturan kategorilerin objektifliğinin, içerik, araştırma süreci, materyal kullanımı, grafik oluşturma, tablo oluşturma sırasıyla arttığı görülmüştür. Bu

bulgudan yola çıkarak, analitik puanlama anahtarı kullanımının puanlayıcılar arasındaki farklılıkları tamamen ortadan kaldırmada yeterli olmadığı; fakat puanlamalar arasında tutarlılığı arttırarak objektiflik düzeyini arttırdığı söylenebilir.

Araştırmada, hem analitik puanlama hem de bütünsel puanlama anahtarından elde edilen puanların üç teknikle yapılan analizlerinde en yüksek güvenirlik değerleri iki puanlayıcı olduğu durumda elde edilmiş, puanlayıcı sayısı arttıkça güvenirliğin giderek düştüğü sonucuna varılmıştır. Bu sonuç Abedi, Baker ve Herl (1995)'in çalışmalarının performansın ölçülmesinde puanlayıcı sayısı artışının puanlardaki değişkenlik düzeyini arttırarak güvenirliği düşürdüğü bulgusuyla ve Nying (2004)'in güvenirliğin puanlayıcı sayısı artışından etkilendiği bulgusuyla örtüşmektedir.

Araştırma kapsamında, Kappa istatistiği tekniğiyle puanlayıcı sayısına dayalı olarak yapılan üç analizde en yüksek uyum iki puanlayıcı olduğu koşulda elde edilmiştir. Kappa istatistiği tekniğinde puanlayıcı sayısının artışı, Kappa değerini düşürmüş; ancak istatistiğin anlamlılık düzeyi aynı kalmıştır. Bu durum Kappa istatistiğinin puanlayıcı sayısından etkilendiğinin göstergesi olabilir. Araştırmanın bu sonucu Nying (2004)'in Kappa istatistiğinin puanlayıcı sayısından olumsuz etkilendiği bulgusuyla paraleldir.

Araştırma sonucunda Krippendorff alfa tekniğiyle yapılan üç analizden en yüksek değerler, iki puanlayıcı arasındaki uyuma ilişkin çıkmıştır. Puanlayıcı sayısının artışı alfa değerini değiştirmiş; ancak bu değişim Kappa istatistiğindeki kadar değişkenlik göstermemiş ve daha kararlı yapı sergilemiştir. Analizde puanlayıcı sayısı ikiden beşe çıkarıldığında alfa değerinde düşüş yaşanmıştır; ancak puanlayıcı sayısı ona çıkarıldığında alfa değerinde önemli düzeyde düşüş olmamıştır. Bu bulgu ışığında, Krippendorff alfa tekniği ile uyum elde edilmek istendiğinde iki ile beş arasında puanlayıcıya yer vermenin yeterli olduğu; beş puanlayıcıdan sonra istatistikte çok değişim olmadığı söylenebilir.

Çalışmada, analitik dereceli puanlama anahtarına dayalı olan log linear analiz sonuçlarına göre puanlayıcılar arasındaki uyumun yanı sıra puanlayıcıların, kategorilerin ve altkategorilerin birbirleriyle etkileşimi de elde edilmiştir. Bu da araştırmacıya daha ayrıntılı bilgi sunmuş ve diğer tekniklerden farklı olarak sonuçlardaki uyumsuzluğun nereden kaynaklandığı konusunda bilgi vermiştir. Analiz sonuçlarının uyumsuzluk hakkında bilgi vermesi puanlayıcılar arasındaki uyumsuzluğa neden olan değişkenlerin araştırmadan çıkarılmasına imkân sağlayabilmektedir.

Araştırma neticesinde, daha detaylı ölçme sonuçları elde edilmek istendiğinde alt kategorilerden oluşan analitik puanlama anahtarı kullanılarak toplanan puanlar, kategorik veri analizi için uygun olan log linear analiz tekniği ile analiz edilebilir. Diğer açıdan daha genel ölçme sonuçlarına ulaşılmak istendiğinde bütünsel puanlama anahtarı ile elde edilen puanların Krippendorff alfa tekniği kullanılarak analiz edilmesinin uygun olduğu düşünülebilir. Sonuç olarak, performansın ölçülmesinde güvenirliğin belirlenmesinde yararlanılacak tekniklerin hangisinin seçileceği, elde edilen puanların hangi amaç doğrultusunda kullanılacağına, puanların hangi ölçek türünde elde edildiğine, analiz sonucunda sağladıkları bilgilerin yapılan ölçme işleminin amacına uygunluğuna bağlı olarak değişmektedir.

## KAYNAKÇA

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

Airasian, P. W. (1994). *Classroom assessment.* New York: McGraw-Hill.

Agresti, A. (1996). *An introduction to categorical data analysis*. New York: John Wiley & Sons, INC.

Agresti, A. & Yang, M. (1987). An empirical investigation of some effects of sparseness in contingency tables. *Computational Statistics & Data Analysis, 5*, 9-21.

Atılgan, H., Kan, A. ve Doğan, N. (2007). *Eğitimde ölçme ve değerlendirme* (2. Basım). Ankara: Anı.

Baykul, Y. (2000). *Eğitim ve psikolojide ölçme: Klasik Test Teorisi ve uygulaması.* Ankara: ÖSYM.

Brennen, R. L., & Prediger, D. J. (1981). Coefficient Kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement, 41*(1981), 687-699.

Burry-Stock, J. A., Shaw, D. G., Laurie, C., & Chissom, B. S. (1996). Rater-agreement indexies for performance assessment. *Educational and Psychological Measurement, 56*(2), 251-262.

Cohen. J. R., Swerdlik M. E., & Phillips, S. M. (1996). *Psychological testing and assessment*. (3th Ed.). London: Mayfield.

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

Crawforth, K. (2001). M*easuring the interrater reliability of a data collection instrument developed to evaluate anesthetic outcomes* (Doctoral Dissertation). Available from Proquest Dissertations and Theses database. (UMI No. 3037063)

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Ohio: Centage Learning.

Fitzpatrick, R., & Morrison, E. J. (1971). Performance and product evaluation. In R. L. Thorndike (Ed.), *Educational measurement* (p. 237–270). Washington DC: American Council on Education.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*(5), 378-382.

Goodrich, H. (1997). Understanding rubric. *Educational Leadership, 54*(4), 14-17.

Goodwin, L. D. (2001). Interrater agreement and reliability. *Measurement in Psychical Education and Exercises Science*, *5*(1), 13-14.

Haladyna, M. T. (1997). *Writing test items to evaluate higher order thinking*. Needham Heights: Allyn and Bacon.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Rewiew*, *2*(2007), 130-144.

Korkmaz, H. (2004). *Fen ve teknoloji eğitiminde alternatif değerlendirme yaklaşımları.* Ankara: Yeryüzü.

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

Krippendorff, K. (2004). Measuring the reliability of qualitative text analysis data. *Humanities, Social Sciences and Law, 38*(6), 787-800.

Krippendorff, K. (2007). Computing Krippendorff's alpha reliability. 7 Eylül 2015 tarihinde http://repository.upenn.edu/asc_papers/43/ adresinden erişildi.

Kutlu, Ö., Doğan, D. C. ve Karakaya, İ. (2009). *Öğrenci başarısının belirlenmesi: performansa ve portfolyaya dayalı durum belirleme.* Ankara: Pegem Akademi.

Landis, J, R., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159-174.

Mertler, C. A. (2001). Designing scoring rubrics for your classroom. *Pratical Assessment Research and Evaluation, 7*(25). Available online: http://PAREonline.net/getvn.asp?v=7&n=25.

Moskal, B. M. (2000). Scoring rubrics: What, when and how? *Practical Assessment Research and Evaluation, 7*(3). Available online: http://PAREonline.net/getvn.asp?v=7&n=3.

Nitko, A. J. (2001). *Educational assessment of students.* (3th ed). New Jersey*:* Prentice Hall.

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

Sim, J., & Wright, C. C. (2005) The Kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Theraphy*, *85*(3), 258-268.

Tanner, M. A., & Young, M. A. (1985). Modeling agreement among raters. *Journal of the American Statistical Association*, *80*(389). 175-180.

Viere, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: The Kappa statistic. *Family Medicine, 37*(5), 360-362.

Von Eye, A., & Mun, E. Y. (2005). *Analyzing rater agreement*: *Manifest variable methods*. New Jersey: Lawrence Erlbaum Associates.

## EXTENDED ABSTRACT

### *Introduction*

The concept of performance based assessment has gained currency in recent years instead of traditional conception of measurement. Analytic and holistic rubrics are used in performance based assessments as the instruments of measurement. The instruments employed in performance assessment should be valid and reliable, as in other instruments of measurement. Reliability is defined as freedom of measurement results from random errors (Baykul, 2000). In the process of performance-based assessment, where raters are the undesired source of variability, techniques for interrater reliability assessment have been recommended. Interrater reliability is defined as no change of rating from one rater to another (Kutlu et al., 2009). It is apparent from the literature that several techniques have been

employed in evaluating interrater reliability (Jonsson & Svingby, 2007). This study aims to analyse and compare the reliability values calculated for the performance prepared for the same purpose and administered to the same individuals by using analytic and holistic rubric through Kappa, Krippendoff, Log-linear analysis techniques. It also aims to find whether rater reliability varies depending on the number of raters and the techniques used.

### *Method*

This study is built on the basis of applying Kappa statistis, Krippendorff Alpha coefficient and log-linear analysis techniques, on determining the similarities as well as differences of these techniques, examining their restrictions, and on finding which of these techniques provide more information. Since it tries to determine the situation, it is a descriptive study.

The study group was composed of 50 students fulfilling performance tasks and reporting them, and 10 teachers rating the tasks by using two rubrics: one of the rubrics was analytic and the other one was holistic. The students in the study group were 50 students who were the $5^{th}$ graders in a state school located in Beyoglu district of Istanbul- where one of the researchers was teaching. Such factors as reaching the students easily and being able to monitor them during application were influential in determining the participants. The teachers in the study group were the elementary school teachers employed in Istanbul in the body of the Ministry of National Education. The participating teachers were chosen on the basis of volunteering.

Analytic and holistic rubrics were used in this study as the tool of data collection. Analytic rubric, developed by Kutlu et al (2009), was composed of five sub-categories-namely, content, the process of research, use of materials, graphic formation, and tabulation. In each sub-category performance was scored between 1 and 3. The maximum score receivable was 15 while the minimum score receivable was 5. Holistic rubric was created on the basis of analytic rubric, which was developed by Kutlu et al (2009) and its final shape was given by consulting five measurement and evaluation experts' opinions. The maximum score receivable from this instrument was 4 whereas the minimum score was 1.

### *Results and Discussion*

Ratings obtained by using analytic rubric yielded relatively more objective results than the one obtained by using holistic rubric- even though both rubrics were designed for the same purpose. Thus, it was concluded that analytic rubric- which produced more standard and more objective results- yielded more consistent results and that therefore it was more reliable. This finding is in parallel to the one obtained by Kutlu et al (2009) stating that the results obtained with the use of holistic rubric were less reliable than those obtained with the use of analytic rubric. The finding is also supportive of Johnsson and Svingby's (2007) conclusion that analytic rubrics increase reliability.

This study analysed the results obtained by using analytic rubric containing sub-categories. It was found that the categories constituting the rubric raised the level of objectivity in the order of content, the process of research, use of materials, graphic formation and tabulation. Based on this finding, it may be said that using analytic rubric is not adequate on its own in eliminating interrater differences completely but that it increases interrater consistency and thus it raises the level of objectivity.

The highest reliability values in analyses performed through all three techniques in the scores obtained from both analytic and holistic rubrics were reached when there were two raters, and it was observed that reliability decreased gradually as the number of raters increased. This finding is in parallel to the Abadi, Baker, Herl's (1995) finding that increase in the number of raters causes a rise in the level of variability in scores and a decrease in reliability in performance measurement, and to Nying's (2004) finding that reliability is influenced by increase in the number of raters.

The highest agreement between raters was observed when there were two raters in all three analyses performed by using Kappa statistics and Krippendorff alpha techniques. The increase in the number of raters reduced the value; however, the significance level of the statistics remained the same. On

examining the Kappa and Alpha results together, it was found that the increase in the number of raters brought about a change in values but that the change varied as much in Alpha results as in Kappa statistics and that it was stable. On raising the number of raters from two to five in analyses, a fall was observed in Alpha values, but when the number of raters was raised to ten, no significant decrease was seen in Alpha values.

The interactions of raters, categories, and sub-categories in addition to interrater agreement were also obtained in log linear analysis results in this study. This provided the researcher with more detailed information, and as different from other techniques, it also informed the researcher about the source of disagreement between analysis results. Presenting information on the disagreement between analysis results enables researchers to remove the variables causing disagreement from their study.

If the intention is to obtain more detailed measurement results from research, scores obtained by using the analytic rubric which is composed of sub-categories can be analysed by using log-linear analysis technique- which is suitable for categorical data analysis. On the other hand, if the purpose is to obtain more general measurement results, analysing the scores obtained through holistic rubric by using the Krippendorff technique would be more appropriate. In consequence, which technique to choose in determining reliability in performance measurement changes depending on for what purpose to use the scores obtained, in what type of scale the scores have been obtained, and on the suitability of the data provided in consequence of the analyses to measurement purpose.