HACETTEPE ÜNİVERSİTESİ
EĞİTİM BİLİMLERİ ENSTİTÜSÜ

Eğitim Bilimleri Ana Bilim Dalı
Eğitimde Ölçme ve Değerlendirme Programı

PUANLAYICI DENEYİMLERİNE GÖRE PUANLAYICILAR ARASI GÜVENİRLİKLERİN FARKLI YÖNTEMLERLE İNCELENMESİ

Mehmet Ali AYDOĞMUŞ

Yüksek Lisans Tezi

Ankara, 2021

Liderlik, araştırma, inovasyon, kaliteli eğitim ve değişim ile

Daha ileriye... En İyiye...

HACETTEPE ÜNİVERSİTESİ
EĞİTİM BİLİMLERİ ENSTİTÜSÜ

Eğitim Bilimleri Ana Bilim Dalı
Eğitimde Ölçme ve Değerlendirme Programı

PUANLAYICI DENEYİMLERİNE GÖRE PUANLAYICILAR ARASI GÜVENİRLİKLERİN FARKLI YÖNTEMLERLE İNCELENMESİ

ANALYSIS OF INTERRATER RELIABILITY WITH DIFFERENT METHODS ACCORDING TO RATER EXPERIENCE

Mehmet Ali AYDOĞMUŞ

Yüksek Lisans Tezi

Ankara, 2021

Öz

Bu araştırmanın amacı, yazma becerilerini ölçen açık uçlu yazma görevlerinin, puanlama sürecine ilişkin deneyimi olan ve deneyimi olmayan öğretmenler tarafından analitik ve bütünsel dereceli puanlama anahtarı yardımıyla puanlanmasından elde edilecek güvenirlik değerlerinin karşılaştırılmasıdır. Farklı tekniklerle belirlenen güvenirlik ve tutarlılık düzeylerinde deneyimli-deneyimsiz gruplar arasında farklılaşma olup olmadığının belirlenmeye çalışıldığı bu araştırma kapsamında MEB Ölçme, Değerlendirme ve Sınav Hizmetleri Genel Müdürlüğünce yürütülen "Yazılı Anlatım Becerilerinin Ölçülmesi ve Değerlendirilmesi" çalışması için uygulanan "Yazma Becerileri Testi" verileri kullanılmıştır. Türkçe yazma becerilerini ölçen ve 8 adet açık uçlu sorudan oluşan bu test Adana, Ankara ve İstanbul illerinde bulunan okullar arasından tabakalı örnekleme yöntemine göre seçilen okullarda 2016-2017 Eğitim-Öğretim yılı bahar döneminde uygulanmıştır. Bu amaçla ilkokul 4, ortaokul 7 ve ortaöğretim 9. sınıflarda öğrenim gören toplam 240 öğrenciye uygulanan "Yazma Becerileri Testi"nden elde edilen sonuçlar puanlama deneyimine sahip olan ve puanlama deneyimi olmayan iki ayrı grup olmak üzere 12 sınıf öğretmeni, 12 Türkçe öğretmeni ve 12 Türk dili ve edebiyat öğretmeni tarafından puanlanmıştır. Puanlama sonuçlarının Kappa tekniği, Krippendorf'un Alpha katsayısı ve Genellenebilirlik Kuramı'na dayalı güvenirlik kestirimleri yapılarak puanlayıcılar arası tutarlılık düzeyine bakılmıştır. Araştırma sonucunda tüm gruplarda Kappa ve Krippendorff'un Alpha tekniğinden elde edilen katsayılarda paralellik ve tutarlılık gözlenmiştir. Yapılan Genellenebilirlik kuramı çalışmalarından elde edilen G ve Phi katsayıları incelendiğinde deneyimsiz Türkçe öğretmenleri ile Türk dili edebiyatı öğretmenlerinin her iki grubunda güvenirlik ve genellenebilirlik için beklenen (0,80) değerine ulaşıldığı gözlenmiştir. Bununla birlikte puanlama deneyiminin tek başına güvenirliğe belirgin bir etkisine dair bulguya ulaşılamamıştır.

Anahtar sözcükler: genellenebilirlik kuramı, puanlayıcılar arası güvenirlik, kappa, krippendorf, yazma becerileri

Abstract

The aim of this study was to compare the reliability values obtained from the scoring of open-ended items measuring writing skills by teachers with and without experience in the scoring process with the use of analytical and holistic rubrics. Within the scope of this research, in which it is tried to determine whether there is a difference between experienced and inexperienced groups in the levels of reliability and consistency determined by different techniques, the "Writing Skills Test" data conducted by the General Directorate of Measurement, Evaluation and Examination Services of the Ministry of National Education were used. For this purpose, the results obtained from the writing test which is applied to a total of 240 students studying in the 4th, 7th and 9th grades were scored by two separate experienced an inexperienced groups. Reliability estimations of the scoring results based on Kappa technique, Krippendorff's Alpha coefficient and Generalizability Theory were made and the level of consistency between raters was investigated. As a result of the research, parallel and consistent values were observed in the coefficients obtained from Kappa and Krippendorff's alpha technique in all groups. When the G and Phi coefficients obtained from the generalizability theory studies were examined, it was observed that the expected (0,80) value for reliability and generalizability was reached in inexperienced Turkish teachers and both groups (experienced and inexperienced) of Turkish language and literature teachers. However, there was no evidence of a significant effect of scoring experience alone on reliability.

Keywords: generalizability theory, inter-rater reliability, kappa, krippendorf, writing skills

Teşekkür

Bu araştırmanın hazırlanma sürecinde her soruma içtenlikle cevap veren, karşılaştığım sorunlarda desteğini esirgemeyen ve elinden gelen tüm yardımı sunan değerli hocam ve akademik danışmanım Prof. Dr. Nuri DOĞAN'a,

Tez savunma jürimdeki yapıcı önerileri ile tezime yaptıkları önemli katkılar için değerli hocalarım Prof. Dr. Selahattin GELBAL ve Dr. Öğr. Üyesi Gökhan AKSU'ya,

Yüksek lisans eğitimim boyunca derslerde öğrettikleri bilgiler ve kazandırdıkları tecrübeler için tüm değerli hocalarıma,

Tez çalışmam için gerekli verileri paylaşarak araştırmamı yürütmemi sağlayan Millî Eğitim Bakanlığı Ölçme, Değerlendirme ve Sınav Hizmetleri Genel Müdürlüğü Araştırma, Geliştirme ve Projeler Daire Başkanlığına ve çok değerli yönetici ve çalışanlarına,

Akademik süreçte her adımda fikirlerine başvurduğum dostlarım Amine CANIDEMİR ve Gülten ŞEN'e,

Bu yolda her koşulda hem maddi hem manevi olarak yanımda olan ismini sayamadığım tüm dostlarıma,

Her daim yanımda olan ve desteklerini hiçbir zaman esirgemeyen Gökçe, Zeynep, Deniz ile sevgili anne ve babama

teşekkürlerimi sunarım.

İçindekiler

Tablolar Dizini

Şekiller Dizini

Simgeler ve Kısaltmalar Dizini

ABİDE: Akademik Becerilerin İzlenmesi ve Değerlendirilmesi

AERA: American Educational Research Association

ANOVA: Analysis of Variance

D: Karar

DPA: Dereceli Puanlama Anahtarı

EFL: English as a Foreign Language

ESL: English as a Second Language

G: Genellenebilirlik

KTK: Klasik Test Kuramı

MEB: Millî Eğitim Bakanlığı

MELAB: Michigan English Language Assessment Battery

ÖDSGM: Ölçme Değerlendirme ve Sınav Hizmetleri Genel Müdürlüğü

PISA: Programme for International Student Assessment

TIMSS: Trends in International Mathematics and Science Study

Bölüm 1
Giriş

Bu çalışmanın yapılma gerekliliği olan problem durumu, araştırmanın amacı ve önemi, problem cümleleri ile sayıltı ve sınırlılıklara bu bölümde yer verilmiştir.

Problem Durumu

Eğitimde başarının ölçülmesi ve değerlendirilmesi eğitim-öğretim sürecinin ayrılmaz bir parçasıdır. Eğitim sürecinde uygulanan öğretim programlarının etkili ve dinamik olması istenir. Eğitim programlarına dinamik bir yapı sağlamak ve yenilenme sürecini devam ettirmek için ölçme ve değerlendirme yöntemlerindeki yenilikler önemli derecede katkılar sağlayabilir. Öğrenme hedeflerine hangi derecede ulaşıldığı da yine ölçme ve değerlendirme süreci ile gözlenebilir. Değerlendirme eğitim sistemini kontrol eden bir görev üstlenir. Değerlendirme, kendisi de dâhil olmak üzere, eğitim sistemindeki öğelerin iyi işleyip işlemediğini, varsa işlemeyen yönlerini ortaya koyar ve eğitim sisteminin eksik yanlarının onarılması için şarttır (Baykul, 2000). Değerlendirme, “ölçme sonuçlarını bir ölçüte vurarak, ölçülen nitelikler hakkında bir değer yargısına varma süreci” olarak tanımlanmaktadır (Turgut, 1997).

Eğitimde kullanılan farklı ölçme araçları vardır. Bunlar açık uçlu görevlerden oluşan testler, kısa cevaplı testler, doğru-yanlış tipi testler, çoktan seçmeli testler, sözlü yoklamalar olarak sınıflandırılabilir. Bu araçlardan hangisinin kullanılacağı yoklanmak istenen özelliğin doğasına göre veya pratik nedenlere göre değişiklik gösterebilir. Açık uçlu görevlerden oluşan testler, az sayıda soru sorularak, öğrencilerin cevaplarını düşünüp istedikleri gibi düzenleyerek oluşturmalarına olanak tanıyan bir sınav türüdür. Öğrencilerin cevaplarını kendilerinin oluşturması puanlamanın objektifliğini etkileyen unsurlardan birisidir. Öğrencilerin bir kelime, bir rakam, bir ibare veya en çok bir cümle ile cevaplayabileceği görevlerden oluşan test tipine kısa cevaplı test denir (Turgut 1997).

Ölçme araçları sonuçlarının değerlendirilmesi bakımından objektif ve subjektif puanlanan olmak üzere ikiye ayrılabilir. Ölçme durumunun amacına göre seçilebilecek olan bu tür araçlar ilgili duruma göre avantaj ve dezavantajlara sahip

olmakla birlikte belki de en önemli ayrımları objektif ve subjektif şekilde puanlanmalarıdır. Söz konusu araçlardan elde edilen puanların güvenirliğini berlirlemede farklı yöntemler bulunmakla birlikte bu çalışmada sübjektif puanlanan araçlardan elde edilen sonnuçların güvenirliğini belirlemeye odaklanıldığı için sübjektif puanlanan araçlar üzerinde durulacaktır. Bu bağlamda özellikle bu çalışmada ele alınan yazma becerilerini ölçmek amacıyla kullanılmış araçlara odaklanılacaktır.

Açık uçlu görevlerden oluşan testler öğretmenler tarafından sıkça kullanılan bir sınav türüdür. Çünkü bazı bilişsel özellikler objektif testlerle yoklanmaya uygun değildir. Yazma becerilerini ölçen açık uçlu görevler ise belirlenmiş net bir doğru cevabının bulunmaması yönüyle açık uçlu soruların diğer türlerinden ayrılır. Yazma becerilerini ölçen açık uçlu sorular öğrencilerin dili kullanması, bilgileri ve düşünceleri organize etmesi; etkili bir anlatımla ifade etmesi gibi özellikleri ölçmeye yöneliktir. Bu nedenle bu tür sorularla ölçülen özelliklerin çoktan seçmeli testlerle yoklanması geçerlik açısından sakıncalar içerebilir. Çünkü yazma becerisini ölçen açık uçlu sorular öğrencinin herhangi bir düşünceyi planlamasını, yapılandırmasını ve kendi kelimeleriyle yazmasını gerektirir.

Yazma becerisini değerlendirmenin bir yolu performans değerlendirmesini kullanmaktır. Performans değerlendirmesi, öğrencinin bir ürün oluşturmasını, bir süreci istenilen kriterler doğrultusunda göstermesini veya her ikisini birden içeren bir tasarımı sunmasını gerektirir (Nitko ve Brookhart, 2016). Bu süreç değerlendirilen becerinin performansının sergilenmesi ve performansın puanlayıcılar tarafından değerlendirilmesi şeklinde ilerler. Dolayısıyla performans değerlendirmeleri, gerçek dünyadaki davranışı temsil etme dereceleri bakımından geleneksel kağıt ve kalem testlerinden farklılık gösterir. Bu anlamda, örneğin çoktan seçmeli maddeleri işaretlemenin aksine gerçek yazma durumunu içeren herhangi bir yazma testi, yazılı bir ürün ortaya koyma performansını temsil ettiğinden, bir performans testi olarak kabul edilebilir.

Yazma becerilerinin ölçülmesinde sıklıkla kullanılan performans değerlendirmelerindeki önemli sınırlılık, puanlamadaki güvenirliğin nispeten düşük olabilmesidir. Birçok ciddi çalışmada bu konu incelenmiş ve görülmüştür ki bu tip sorular farklı öğretmenler tarafından farklı şekilde puanlanmıştır. Hatta bu tip görevlerin aynı öğretmenler tarafından farklı zamanlarda farklı puanlandığı da

görülmüştür. Diğer bir sınırlılık ise puanlama için gereken yeterli süredir. Eğer puanlayıcı dikkatli bir şekilde puanlama yapar ve gerekli notları yeri ve zamanı geldiğinde almaya özen gösterirse az sayıdaki sınav kâğıtlarını puanlamak için bile saatler harcamak gerekecektir (Gronlund 1998).

Yazılı anlatım becerilerini ölçmek için en uygun yöntemlerden biri açık uçlu görevlerin kullanılmasıdır. Objektif çoktan seçmeli testler kullanmak yazma becerisinin ölçülmesinde yeterli olmaz ve geçerlik sorunu ortaya çıkarabilir. Yazılı anlatım becerilerini ölçmeye yönelik olarak yapılan bu tür sınavların güvenirliği ve geçerliği önemli bir konudur. Yazma işi öznel öğeleri fazlaca içerdiğinden dolayı, güvenilir olarak puanlanması oldukça güçtür. Bu tür sınavlarda net ve kesin bir cevap anahtarı hazırlamak mümkün değildir. Bunun yerine içerik, organizasyon, dilbilgisi, anlatım yapısı, noktalama ve yazım kuralları gibi ana başlıklar altında uygunluk ve yaratıcılıkları kapsayan bir cevap anahtarı hazırlanır. Dolayısıyla bu gibi özellikleri puanlarken objektiflik güçleşir.

Ölçmenin hata kaynaklarını belirleme ve puanların güvenirliğinin kestirilmesi işlemi, yazma becerilerinin değerlendirilmesi de dâhil olmak üzere herhangi bir ölçme sonucunda elde edilen sonuçların doğru yorumlanması için esastır (AERA, APA, NCME, 1999). Yazma ve diğer benzer performans değerlendirmelerinde, puanlayıcılar ve görevler oluşabilecek potansiyel ölçme hatalarının genellikle birincil kaynaklarıdır. Görev örneklemine ilişkin değişkenliğin ölçme hatasının ana kaynağı olduğunu gösteren önemli göstergelerle karşılaşılabilmektedir. Bununla birlikte, puanlayıcı farklılıkları, iyi eğitimli puanlayıcıların bulunduğu ve iyi tasarlanmış ölçme tasarımları için puanların değişkenliğine önemli ölçüde etki etmez (Gao, Brennan ve Guo, 2005). Dolayısıyla iyi bir ölçme tasarımı oluşturmanın yanında, puanlayıcıların yapılan işe hâkim ve bu alanda eğitim almış olmalarının sonuçların güvenirliği açısından önemli olduğu söylenebilir. Yazma becerilerinin değerlendirilmesi sınıf-içi çalışmalarda da önemli ölçüde yer alabilir. Öğretmenler öğrencilerinin iyi yapılandırılmış ve anlaşılır metinler yazıp yazamayacağını bilmek isteyebilirler. Tüm performans değerlendirmelerinde olduğu gibi, yazma yeterliklerini değerlendirmenin de en uygun yolu, bireylerden bir veya daha fazla metin yazmasının istenmesidir (Huot, 1990b). Bununla birlikte, bu değerlendirmelerin sonucunu yazma becerilerine genellemek zordur. Bunun sebebi bireylerin ortaya koyduğu ürünün kalitesinin yalnızca bireylerin yazma becerilerine bağlı olarak değil, aynı zamanda

puanlayıcılar ve görevler gibi ölçme durumunun özellikleri nedeniyle de farklılık göstermesidir (Bouwer, Béguin, Sanders ve van den Bergh, 2015). Bununla birlikte puanlama süreçleri büyük önem taşımaktadır çünkü bu süreç sonunda elde edilen puan bireyler hakkında karar verirken kullanılacak tek göstergedir.

Pek çok eğitimsel ve psikolojik araştırma, davranışın bazı yönlerini ölçmek için bağımsız puanlayıcıların kullanılmasını gerektirir. Örneğin, puanlayıcılar, standartlaştırılmış bir testte açık uçlu görevleri puanlamak, bir spor müsabakasında uzman sporcuların performansını derecelendirmek veya yeni bir puanlama anahtarının uygulanabilirliğini deneysel olarak test etmek için kullanılabilir. Bu tür puanlama yöntemleri genellikle, ölçme konusu olan davranışların basit bir doğru-yanlış test maddesi gibi objektif olarak puanlama yapılamayan durumlarda kullanılır. Davranışı puanlama görevi puanlayıcının ölçülmek istenen yapıya ilişkin yorumuna bağlı olacağı için bir dereceye kadar yanlılığı da içinde barındırır. Bu yanlılığı azaltmak için bir strateji, puanlama anahtarları geliştirmektir (White, 1984).

Puanlayıcı güvenirliği iki veya daha fazla puanlayıcı arasında bulunan uyumun veya tutarlılığın bir derecesidir (Crocker ve Algina, 1986). Puanlayıcıların yer aldığı ölçme durumlarında, puanlayıcılar arası güvenirliğin derecesini kestirmek önemlidir. Çünkü bu derece araştırma sonuçlarının geçerliği ve güvenirliği için önemli çıkarımlara sahiptir. Örneğin iki puanlayıcının, bireyleri gözlemlenen davranışlara göre güvenilir bir şekilde puanladığı sonucuna ulaşılamazsa, bu puanlayıcılar tarafından verilen puanların analizleri doğru olmayan sonuçlar verecektir. Ayrıca, bir araştırmada geçmişte güvenirliği kanıtlanmış puanlama anahtarları kullanılıyor olsa dahi, her yeni araştırma için puanlayıcılar arası güvenirlik yeniden gösterilmelidir. Puanlayıcılar arası güvenirlik, bir puanlayıcı grubu arasında belirli bir zamanda aynı yanıtlar üzerindeki anlaşma düzeyinin bir ölçüsüdür. Bu nedenle, puanlayıcı güvenirliği ölçme aracının değil, test durumunun bir özelliğidir. Puanlayıcılar arası güvenirlik çalışmaları yapmadan önce şu üç soruyu sormak faydalı olacaktır (Stemler ve Tsai, 2008):

Puanlayıcılar arası güvenirlik çalışması yapılmasının amacı nedir?

Veriler hangi türdedir?

Çalışmayı yürütmek için hangi kaynaklar mevcuttur? (teknik uzmanlık, zaman, maddi olanaklar vb.)

Geniş ölçekli sınavlarda puanlayıcılar arası yüksek düzeyde güvenirliğe ulaşmak için White (1984) tarafından bir dizi prosedür önerilmiştir. Bunlardan ilki puanlama anahtarları kullanmaktır. Ölçütlerin belirlendiği ve sınırların çizildiği bir puanlama anahtarıyla daha uyumlu ve güvenilir sonuçlara ulaşmak mümkündür. Bir diğer yöntem, puanlayıcılar için örnek-istenen yanıtlar oluşturmaktır. Bu sayede puanlayıcıların zihninde somut bir gösterge oluşabilir ve puanlama daha net olarak yapılabilir. Her yanıt en az iki puanlayıcı tarafından puanlanmalıdır. Bu puanlayıcılar arasında uyuşmazlık olduğu durumlarda ilgili yanıt bir üçüncü “üst” değerlendirici tarafından yeniden incelenmelidir. Puanlama işlemi puanlayıcıların bir araya gelerek kontrollü bir puanlama sürecini gerçekleştirmesiyle yapılabilir. Böyle bir yöntemin kullanılmasının avantajı, kaynağı belli olmayan hata varyanslarını en aza indirmesidir. Ancak bu işlem dikkatli bir şekilde yapılmadığında puanlayıcıların aynı puanları verme eğilimi gibi bir yanlılık hatasına düşmeleri söz konusu olabilir. Puanlama, alanında uzmanlaşmış bir “lider puanlayıcı” tarafından diğer puanlayıcıların kontrol edilmesi şeklinde olabilir. Bu sayede puanlayıcıların standartlara uygun davranıp davranmadığı süreç içerisinde gözlenebilir. Son olarak puanlama sürecinin gerçek zamanlı izlenmesi büyük fayda sağlayabilir. Bu sayede tutarlılık gösteren puanlayıcılar sürece devam edebilecekken, tutarsızlık oranı belirli bir düzeyin altında kalan puanlayıcılar sistemden çıkarılabilir.

Stemler ve Tsai (2008)’e göre puanlayıcılar arası güvenirlik çalışması yapmak için üç ana sebep bulunmaktadır. Bu sebeplerden en yaygın olanı, araştırmacının veri analizi ve istatistiksel modellemede kullanılmak üzere bir son puana ulaşmak istemesi ancak öncelikle puanlamanın "öznel" veya " önyargılı" olup olmadığını belirlemesi gerektiğidir. Örneğin, geniş ölçekli ulusal ölçme uygulamalarında, her öğrencinin mevcut akademik başarı düzeyinin genel bir değerlendirmesini sağlamak amacıyla yazma görevlerini puanlamak için birden çok puanlayıcı kullanılması yönteminde amaç budur. Bu gibi durumlarda, puanlayıcılar arası güvenirliğin rapor edilmesi genellikle sadece amaca yönelik bir araçtır ve araştırmacılar bu tür bir güvenirlik analizinin ayrıntılarına inmeyebilirler. Puanlayıcılar arası güvenirlik çalışması yapmanın ikinci sebebi, yeni geliştirilen bir rubriğin amaca hizmet edip etmediğini veya değiştirilmesi gerekip gerekmediğini görmek için değerlendirmektir. Bir puanlama sürecinde puanlayıcılar ne kadar deneyimli ve alan uzmanı olursa olsun, kullanılan puanlama anahtarı istenilen

özellikte değilse sonuçların güvenilir olduğundan söz etmek güçtür. Son olarak, puanlayıcılar arası güvenirlik çalışması yürütmenin üçüncü bir nedeni, puanlamaların bilinen bir "gerçek" durumu ne kadar iyi yansıttığını doğrulamaktır. Geçerlik çalışmaları buna örnek olabilir. Sonuç olarak puanlama sürecinin doğrudan etkileyeceği kararların güvenirliği açısından bu amaçlardan herhangi birisi doğrultusunda araştırma yürütülebilir.

Puanlama süreçlerinde uyumsuzluk yaşanması kaçınılmazdır ve puanlayıcılar hata kaynaklarından birisidir. Aynı yazma görevinin iki farklı puanlayıcı tarafından, hatta bazen aynı puanlayıcı tarafından farklı zamanlarda puanlanması arasında büyük farklılaşmalar olabileceği birçok kez belirtilmiştir (Akt. Cooper, 1984). Puanlayıcılar değerlendirme yaparken her zaman aynı fikirde olmazlar ve çoğu zaman da aynı fikirde değildirler (Schoonen, Vergeer ve Eiting, 1997). Puanlama tutarsızlıkları yazma görevi daha fazla yanıt özgürlüğüne imkân sağladığında, yani istenilen yanıtın sınırlarının net olarak belirlenmediği durumlarda artma eğilimi göstermektedir (Coffman, 1971). Bilgili ve deneyimli bir puanlayıcının bile, yazma görevlerini değerlendirmek için kullanılan kriterlere verdiği nispi ağırlıklar bir görevden diğerine farklılık göstermektedir. Puanlayıcıların, iyi verilmiş yanıtlara kıyasla, yetersiz yanıtlarda hataları arama ve vurgulama olasılığı çok daha yüksek olmaktadır. Bu hale etkisi tam tersi şekilde de meydana gelebilir (Cooper, 1984). Özellikle kontrol edilmeyen süreçlerde gerçekleştirilen puanlama çalışmalarında, örneğin yorgunluk sebebiyle, puanlamada tutarsızlıklar artabilmektedir. Bu durum puanlayıcı yorgunluğu olarak ifade edilmektedir. Okuyucu yorgunluğu önemli bir hata kaynağıdır. Puanlama süreci uzadıkça puanlayıcılar daha düşük puan verme eğiliminde olmaktadır. Bununla birlikte bu yorgunluğun hangi noktada puanlamaya etki etmeye başladığı tam olarak bilinmemektedir (McColly, 1970). Bu uyuşmazlıkları gidermenin bir yolu kontrollü bir puanlama süreci gerçekleştirmektir. Weigle (1994)'ye göre, puanlayıcıların eğitilmesi de bu uyuşmazlıkları azaltmaya yönelik bir seçenektir. Dereceli Puanlama Anahtarı'nın (DPA) kullanımıyla bağlantılı olarak puanlayıcı eğitiminin hedeflenen puanlama kriterlerini netleştirdiği ve puanlayıcıları kendilerininkinden ziyade hedeflenen kriterlere göre değerlendirmeye teşvik ettiği varsayılmaktadır (Charney, 1984). Eğitim aynı zamanda farklı seviyedeki puanlayıcıları eşitleme amacı güderek uyuşmazlığı azaltmayı hedefler ve puanlayıcıların uygun kritere odaklanmasına

yardımcı olur. Puanlayıcıların eğitiminde unutulmaması gereken bir nokta, puanlayıcıların uyum sağlamaları için eğitilmesinin, onları görevleri değerlendirmede kendi deneyimlerini ve uzmanlıklarını görmezden gelmeye zorladığı ve böylece bir metnin birden fazla doğru cevabı olma olasılığını inkâr edebileceğidir (Barritt, Stock ve Clark, 1986). Bu eğitimler, puanlayıcıların olumsuz etkilerini azaltmak amacıyla, onlardan istenen puanlama davranışını netleştirmek ve belirli seviyelerde örnek yazma performanslarını göstermek için tasarlanmaktadır.

Puanlayıcıların eğitiminde literatürde farklı yöntemler bulunmaktadır. Bunlardan biri, puanlayıcıların sözlü protokollerinin (varbal protocols) analizidir. Bu yöntemde puanlayıcılar puanlama işlemini gerçekleştirirken düşüncelerini sözlü bir şekilde ifade ederler. Bu düşüncelerin analizinden puanlama davranışları hakkında yargıya varılmaya çalışılır. Bu süreç daha sonra literatürde “think-aloud procedures (TAP)” yani sesli düşünme yöntemi olarak geliştirilmiştir (Weigle, 1994). Hamp-Lyons (1990), eğitimin gerçekleştiği bağlam, verilen eğitimin türü, eğitimin ne ölçüde takip edildiği, puanlamanın ne ölçüde takip edildiği ve okuyuculara verilen geribildirimler hepsi puanlamanın hem güvenirliğini hem de geçerliğini korumada önemli bir rol oynadığını belirtmiştir.

Günümüzde, geniş ölçekli hizmet veren bazı test geliştirme kuruluşları, puanlama alanında insan yerine bilgisayar yazılımları geliştirmekte ve kullanmaktadır. Bu yazılımlar büyük oranda doğru puanlama yapabilmekle birlikte, halen tam olarak bu sisteme geçilmiş değildir. Dolayısıyla gerçek puanlayıcıların eğitimine yönelik araştırmalar da devam etmektedir.

Puanlayıcı davranışlarındaki bir diğer önemli etken de deneyimdir. Burada bahsedilen deneyimin öğretmenlikten ziyade puanlama çalışmalarına katılma veya bu alanda eğitim almış olma durumu olduğu unutulmamalıdır. Puanlama sürecinde veya öncesinde gerçekleştirilen eğitimlerin puanlama davranışına olumlu etki edebileceği düşünülebilir. Puanlama anahtarlarının nasıl kullanılacağı, görevlerin hangi yönlerine odaklanılacağı, puanlama yapılırken oluşabilecek yanlılıkların önüne nasıl geçilebileceği gibi birçok konuda verilecek eğitimle puanlayıcılar hem güven kazanır hem de yapılacak işin doğasını anlarlar. Bununla birlikte, çoğu alanda olduğu gibi puanlama sürecinde de daha önceden deneyim sahibi olmanın puanlama davranışına birçok yönden etkisi olmaktadır. Daha önce birçok öğrenci yanıtıyla karşılaşmış ve bunları puanlamış bir puanlayıcı, yeni karşılaştığı yanıtları

hangi çerçevede değerlendireceği konusunda daha çabuk organizasyon kurabilir. Öğretim deneyimi olan puanlayıcılar, öğrencileri öğrenmeye teşvik etmenin bir yolu olarak geribildirim vermeye alışkındır (Prior, 1995). Dolayısıyla puanlayıcıların önceki öğretmenlik deneyimi de belirli bazı endişelerini puanlama sürecine taşımaya yatkın olmalarına sebep olabilir (Erdosy, 2003). Örneğin, Wolfe ve Feltovich (1994) yaptığı çalışmada ek puanlama çalışmaları yapıldığında deneyimsiz puanlayıcıların uzman sayılabilecek puanlayıcılarla benzer metin özelliklerine odaklanabileceklerini değerlendirmişlerdir. Freedman (1979), yaptığı çalışmada, üniversite öğrencilerinin yazdığı bazı makaleler içerik, organizasyon, cümle yapısı ve teknik olmak üzere dört kategoride 12 puanlayıcı tarafından değerlendirilmiştir. Bu çalışmanın bir bulgusu eğitimin puanlamanın odağını etkilediğidir. Bu ise puanlayıcıların önceki deneyiminin, yazma becerilerinin değerlendirmesinde yapısal-ilgisiz varyansın (construct-irrelevant variance) bir kaynağı olabileceğidir (Wolfe ve Feltovich, 1994). Cumming (1990), deneyimli öğretmenlerin, ESL (English as a Second Language) makalelerini değerlendirmek için çok sayıda ve çeşitli ölçütler ve bilgi kaynakları kullandığını belirtirken, acemi öğretmenlerin makaleleri genel okuma becerilerinden veya daha önce edindikleri diğer bilgi ve becerileri kullanarak bu kriterlerden yalnızca birkaçıyla değerlendirme eğiliminde olduğuna işaret etmiştir. Cumming'in aynı çalışmadaki bir diğer bulgusu, deneyimli puanlayıcıların yüzeysel özelliklerden daha az etkilenme eğiliminde olduklarını ve aynı anda dil kullanımını, içeriği ve organizasyonu inceleme konusunda daha yetenekli olduklarıdır. Şahan ve Razı (2020), yazma ürününün kalitesinin, puanlayıcıların karar verme davranışları üzerinde puanlama deneyiminden daha büyük bir etkiye sahip olduğunu bulmuşlardır. Barnwell (1989), deneyimli puanlayıcıların dil hatalarına karşı daha hoşgörülü olduğunu tespit etmiş ve puanlayıcıların mümkün olan en geniş dil becerilerine maruz kalmalarının, öğrencileri daha gerçekçi, dolayısıyla daha bağışlayıcı bir bakış açısıyla değerlendirmelerine sebep olduğunu öne sürmüştür.

Genel olarak, puanlayıcılar, belirli bir puanlama anahtarı kullanmaları istendiğinde dahi, puanlama durumlarına kişisel deneyimlerini katabilecekleri gibi, kendilerine rehberlik edecek puanlama anahtarının olmadığı bir durumda daha önceden kullandıkları ölçeklerdeki bilgilerine güvenebilirler (Erdosy, 2003). Puanlayıcı deneyiminin ve puanlama yönteminin yazma puanlarının değişkenliği üzerindeki etkisinin TAP aracılığıyla etkileşimlerinin incelendiği bir araştırmada

Barkaoui (2010), puanlama ölçeğinin türünün, puanlayıcıların karar verme davranışları üzerinde deneyimden daha büyük bir etkiye sahip olduğunu ortaya koymuştur. (Akt. Şahan ve Razı, 2020). Yukarıdaki araştırmalar incelendiğinde, genel olarak puanlama deneyiminin puanlama davranışlarına olumlu veya olumsuz bir etkiden söz etmek yerine değerlendirmenin farklı yönlerine farklı katkıları olduğu söylenebilir. Dolayısıyla puanlama deneyimi ve eğitiminin puanlama davranışı üzerinde bir etkisi olduğundan bahsetmek mümkündür.

Puanlama sonucunda elde edilen verilerin güvenirliğini değerlendirmek için birçok yöntem mevcuttur. Bu bağlamda çatı olarak iki farklı yönden bahsetmek mümkündür. Bunlar puanlayıcı-içi güvenirlik ve puanlayıcılar arası güvenirliktir. Puanlayıcı-içi güvenirlik, bir puanlayıcının farklı durumlarda aynı ürüne aynı puanı verme eğilimini ifade ederken, puanlayıcılar arası güvenirlik, farklı puanlayıcıların aynı ürüne aynı puanları verme eğilimini ifade eder. Bu puanlamalar, en basit anlamda, verilen puanlar arasındaki korelasyon katsayısı aracılığıyla hesaplanabilir. Bu istatistik, iki puan seti arasındaki ilişkinin gücünü gösteren 0 ile 1 arasında bir değer alır. 0'a yakın bir korelasyon katsayısı, birinci puanlayıcı tarafından verilen puanlarla ikincisi tarafından verilen puanlar arasında çok az veya hiç ilişki olmadığını gösterirken, 1'e yakın bir katsayı, puan setleri arasında güçlü bir ilişki olduğunu gösterir. Puanlamaların nasıl analiz edilmesi gerektiği, büyük ölçüde temsil ettikleri veri türüne ve analizin nihai hedeflerine bağlıdır (Gwet, 2014). Puanlayıcılar arası güvenirliği araştırmaya yönelik tamamlayıcı bir yaklaşım, özellikle ikiden fazla değerlendirici söz konusu olduğunda, varyans analizi (ANOVA) yoluyla yapılmaktadır. ANOVA, puanlayıcıların ortalama puanları arasında herhangi bir istatistiksel fark olup olmadığını belirlemek için kullanılabilir (Weigle, 2002).

Puanlayıcılar arası güvenirliklerin belirlenmesi eğitimli veya eğitimsiz puanlayıcılar tarafından gerçekleştirilen puanlamalar ile geniş ölçekli sınavlar veya bireylere yapıcı dönütlerin verilmesinin gerekli olduğu ölçme tasarımları için kullanılabilir. Bununla birlikte, birçok çalışmanın puanlayıcılar arası güvenirliği hesaplarken puanlayıcıların deneyim yıllarını dikkate aldığı ancak daha önce puanlama deneyimi olup olmadığını değerlendirmediği söylenebilir.

Puanlayıcılar arası güvenirliğin incelendiği çalışmalarda bazı yaygın hataların kullanıldığı görülmektedir (Hallgreen, 2012). Bu duruma bir örnek olarak uyum yüzdelerinin kullanılması gösterilebilir. Araştırmalarda rapor edilen uyum yüzdeleri

her ne kadar uyumun derecesi olsa da şansa bağlı uyumu rapor etmediği ve bu ayrıma dair bir kanıt sunmadığı için bu tür değerlendirmelerde kullanılması objektif sonuçlara ulaşılmasını engelleyebilir. Bir diğer hata ise çalışma tasarımına uygun istatistiksel modelin kullanılmaması olarak gösterilebilir. Araştırmada kullanılan verilerin ölçek düzeylerine göre uygun istatistiğin kullanılması doğru sonuçlara ulaşılabilmesine yardımcı olabilir.

Geniş ölçekli sınavlarda çok sayıda bireye uygulanan sorular çok farklı puanlayıcılar tarafından puanlandığı için söz konusu puanlayıcıların puanlama konusunda eğitim almamış olabileceği ya da hepsinin puanlama eğitimi almış olduğu durumda da puanlama deneyimi olanlar ile puanlama deneyimi olmayanlar bu eğitimden farklı etkilenebilir. Bu nedenle hem deneyim hem de eğitim puanlayıcı güvenirliğinde önemli değişkeneler olarak ele alınabilir ve bunları birlikte ele alan çalışmaların gerekli olduğu sonucuna ulaşılabilir.

Alanda yapılmış çalışmalar incelendiğinde, çeşitli kuramlar ve yöntemlerle güvenirlik analizlerinin sıklıkla yapıldığı görülmektedir. Hem sınıf içi değerlendirme uygulamalarında hem de geniş ölçekli sınavlarda yanıtın sınavı alan bireyler tarafından oluşturulduğu ölçme araçları kullanıldığında karşılaşılacak objektiflik sorununu belirlemek amacıyla yapılan bu analizlerde birçok faktör etkili olabilirken bunlardan önemli birisi puanlayıcıların puanlamaya ilişkin deneyim durumları olduğu düşünülebilir. Ancak, puanlayıcıların açık uçlu soruları puanlama deneyimlerine göre puanlama sürecinin güvenirliğine yönelik karşılaştırmaların yapıldığı bir G kuramı çalışması veya puanlayıcıların daha önceki puanlama deneyimlerinin çeşitli yöntemlerle elde edilen güvenirliklerine olan etkisini inceleyen bir çalışmaya rastlanmamıştır.

Araştırmanın Amacı ve Önemi

Bu araştırmanın amacı, MEB ÖDSGM tarafından 2017-2018 yılları arasında uygulanan Yazılı Anlatım Becerilerinin Ölçülmesi ve Değerlendirilmesi Çalışması kapsamında yapılan Yazma Becerileri Testi'nin puanlama çalışmasından elde edilen sonuçlar kullanılarak puanlayıcıların benzer puanlama süreçlerindeki deneyimlerinin puanlamaya etkisini ortaya koymak amacıyla, yazma becerilerini ölçen yazma görevlerinin, puanlama sürecine ilişkin deneyimi olan ve deneyimi olmayan puanlayıcılar tarafından analitik ve bütünsel dereceli puanlama anahtarları

yardımıyla puanlanmasından elde edilecek güvenirlik değerlerinin farklı analiz yöntemleriyle karşılaştırılmasıdır.

Türkiye'de uygulanan geniş ölçekli sınavlarda açık uçlu sorulara geçişin tartışıldığı şu günlerde paydaşların kaygıları özellikle bu tür maddelerin objektif puanlanamaması yönündedir. Geniş ölçekli sınav uygulamalarının sonuçları, birey hakkında sonraki eğitim hayatını etkileyecek önemli kararlar verilmesi amacıyla kullanılmaktadır. Böylesine önemli kararların verildiği durumlarda, özellikle öznelliğe açık olan açık uçlu görevlerin de bulunması, bu sınav süreçleriyle ilgili ek çalışmalar yapılması sonucunu doğurmaktadır. Puanlayıcıların sahip olduğu bireysel farklılıklar, puanlama davranışı üzerinde etkiye sahip olabilir. Bu süreçte dikkate alınabilecek puanlayıcı özelliklerinin başında, puanlayıcıların dereceli puanlama anahtarı kullanmaya ilişkin deneyimleri olabilir. Pratik nedenlerden dolayı deneyim açısından aynı özelliklere sahip puanlayıcılara ulaşmak mümkün olamayabilir. Dolayısıyla, dereceli puanlama anahtarı kullanarak açık uçlu soruları puanlamaya ilişkin deneyimli ve deneyimsiz olan puanlayıcıların, aynı grubu puanlaması kaçınılmazdır. Ancak nitelikli bir puanlama anahtarının, farklı deneyime sahip puanlayıcıların benzer puanlama yapmasına katkı sağlaması beklenir.

Literatür incelendiğinde, puanlayıcı güvenirliği çalışmaları için uyum yüzdesi, Pearson korelasyon katsayısı, sınıf-içi korelasyon katsayısı, Cohen'in Kappa istatistiği, Fleiss'in kappa istatistiği, Gwet'in AC2 katasyısı, Krippendorff'un Alpha katsayısı, ve hata kaynaklarını belirlemeye yarayan genellenebilirlik kuramı ve çok yüzeyli Rasch modeli gibi çeşitli yöntemler mevcuttur. Bu çalışmada, puanlayıcıların deneyimlerinin puanlama güvenirliğine etkileri olup olmadığını belirlemek amacıyla, verilerin de uyum gösterdiği Cohen'in ağırlıklandırılmış Kappa istatistiği, Krippendorff'un Alpha katsayısı ve genellenebilirlik kuramına dayalı hata kaynakları, puanlayıcı bağlamında incelenecektir.

Puanlayıcıların benzer süreçlerdeki puanlama deneyimlerinin puanlama güvenirliğine etkisinin ortaya konacağı bu araştırmanın sonuçlarının puanlayıcılara verilen eğitimlere yönelik programlara ve diğer araştırmalara yol gösterici olması hedeflenmektedir.

Araştırma Problemi

Analitik ve bütünsel dereceli puanlama anahtarı kullanarak açık uçlu yazma görevlerini puanlayan öğretmenlerin benzer süreçlerdeki puanlama deneyimlerine göre puanlayıcı güvenirlik düzeyleri nelerdir? Bu güvenirlik düzeyleri arasında bir ilişki var mıdır?

Alt problemler.

1. Puanlama deneyimine sahip ve daha önce puanlama deneyimi olmayan sınıf öğretmenlerinin puanlayıcılar arası uyum düzeyi nedir?
    a. Cohen'in Ağırlıklandırılmış kappa istatistiğine puanlama deneyimine sahip ve daha önce puanlama deneyimi olmayan sınıf öğretmenlerinin puanlayıcılar arası uyum düzeyi nedir?
    b. Krippendorff'un Alpha katsayısı istatistiğine göre puanlama deneyimine sahip ve daha önce puanlama deneyimi olmayan sınıf öğretmenlerinin puanlayıcılar arası uyum düzeyi nedir?
    c. G kuramına göre puanlama deneyimine sahip ve daha önce puanlama deneyimi olmayan sınıf öğretmenlerinin G ve phi katsayısı düzeyleri nedir?
2. Puanlama deneyimine sahip ve daha önce puanlama deneyimi olmayan Türkçe öğretmenlerinin puanlayıcılar arası uyum düzeyi nedir?
    a. Cohen'in Ağırlıklandırılmış kappa istatistiğine göre puanlama deneyimine sahip ve daha önce puanlama deneyimi olmayan Türkçe öğretmenlerinin puanlayıcılar arası uyum düzeyi nedir?
    b. Krippendorff'un Alpha katsayısı istatistiğine göre puanlama deneyimine sahip ve daha önce puanlama deneyimi olmayan Türkçe öğretmenlerinin puanlayıcılar arası uyum düzeyi nedir?
    c. G kuramına göre puanlama deneyimine sahip ve daha önce puanlama deneyimi olmayan Türkçe öğretmenlerinin G ve phi katsayısı düzeyleri nedir?

3. Puanlama deneyimine sahip ve daha önce puanlama deneyimi olmayan Türk dili ve edebiyatı öğretmenlerinin puanlayıcılar arası uyum düzeyi nedir?
    a. Cohen'in Ağırlıklandırılmış kappa istatistiğine göre puanlama deneyimine sahip ve daha önce puanlama deneyimi olmayan Türk dili ve edebiyatı öğretmenlerinin puanlayıcılar arası uyum düzeyi nedir?
    b. Krippendorff'un Alpha katsayısı istatistiğine göre puanlama deneyimine sahip ve daha önce puanlama deneyimi olmayan Türk dili ve edebiyatı öğretmenlerinin puanlayıcılar arası uyum düzeyi nedir?
    c. G kuramına göre puanlama deneyimine sahip ve daha önce puanlama deneyimi olmayan Türk dili ve edebiyatı öğretmenlerinin G ve phi katsayısı düzeyleri nedir?

Sayıltılar

Farklı öğretmenlerin yaptıkları puanlamaların birbirlerinden bağımsız olarak gerçekleştirildiği varsayılmıştır.

Sınırlılıklar

Bu araştırma 2017-2018 eğitim öğretim yılında Türkiye'de çeşitli illerde görev yapan;

- 12 sınıf öğretmeni, 12 Türkçe öğretmeni ve 12 Türk dili ve edebiyatı öğretmeni ile,
- MEB Ölçme değerlendirme ve Sınav Hizmetleri Genel Müdürlüğü tarafından geliştirilen ve yazma becerilerinin ölçülmesine yönelik hazırlanan 8 açık uçlu yazma görevinden oluşan "Yazma Becerileri Testi"nde kullanılan görevler ile,
- Deneyimli ve deneyimsiz olmak üzere bu çalışmada görev alan öğretmenler ile,

- Ankara, Adana ve İstanbul illerinde okuyan 80 4. sınıf, 80 7. Sınıf ve 80 9. Sınıf öğrencisi olmak üzere toplamda 240 öğrenciden oluşan cevapların yer aldığı örneklem grubu ile sınırlıdır.

Tanımlar

Deneyimli ve deneyimsiz puanlayıcılar: Araştırma kapsamındaki puanlama işlemini gerçekleştiren puanlayıcılar iki farklı başlık altında incelenmiştir. Bunlardan ilki olan deneyimli puanlayıcılar, daha önce en az bir defa yazma görevlerinin puanlanması sürecinde yer almış olan öğretmenleri ifade etmektedir. Deneyimsiz puanlayıcılar ise daha önce herhangi bir yazma görevi puanlama sürecine katılmayan ve bu tür bir süreçte ilk defa puanlama yapacak olan öğretmenleri temsil etmektedir. Sonuç olarak bakıldığında deneyimsiz puanlayıcılar meslek hayatları boyunca sınıf içi değerlendirme süreçleri dışında merkezi olarak bu araştırma kapsamındaki bir sürece ilk defa dâhil olmaları açısından deneyimsiz kabul edilmiştir.

Bölüm 2
Araştırmanın Kuramsal Temeli ve İlgili Araştırmalar

Araştırmanın Kuramsal Temeli

Bu kısımda araştırmaya temel oluşturan puanlayıcı güvenirliği kavramı, puanlayıcı güvenirliğini etkileyen durumlar ve bazı çeşitleri, puanlayıcı güvenirliğini kestirmek için kullanılan KTK'ya dayalı Kohen'in Kappa, Krippendorff'un Alpha teknikleri ve Genellenebilirlik Kuramı detaylı bir şekilde açıklanacaktır. Bununla birlikte Genellenebilirlik kuramı ağırlıkta olmak üzere ilgili yöntemlerin hesaplanması için gereken teknikler ve bu tekniklerin bazı istatistik temelleri de dikkatle bu bölümde ele alınacaktır. İlgili kuram ve tekniklerin güçlü yönleri ile sınırlılıkları da yine bu bölümde ele alınacaktır.

Puanlayıcılar arası güvenirlik. Güvenirlik, genel bir tanımla, bir test durumunun farklı özellikleri veya yönleri arasında ölçüm tutarlılığı olarak tanımlanabilir. Örneğin, bireyler bir puanlayıcıdan diğerine aynı puanı alıyorsa ve bir grup sınav aynı şekilde farklı durumlarda, bir testin farklı sürümlerinde veya farklı puanlayıcılar tarafından aynı şekilde sıralandığında bir testin güvenilir olduğu söylenebilir. Güvenirlik, testlerden elde edilen ölçümlerde bulunması istenen önemli bir husustur ve testlerden elde edilen sonuçların geçerliği için de bir ön koşul niteliğindedir. Yani, bir testin tutarlı sonuçlar verdiğinden emin olunmazsa, test sonuçlarına dayanarak varılan çıkarımların ve kararların tutarlı olduğundan ve gözlenmek istenen yeteneğin yansıtıldığından emin olunamaz. Bu bağlamda, bir yazma görevinden elde edilen sonuçların güvenirliği, yazma görevinin kendisiyle ilgili olabileceği gibi puanlama süreciyle ilgili değişkenlerden de etkilenir. Örneğin puanlayıcıların kendi alanlarındaki geçmişi ve deneyimi, dereceli puanlama anahtarının yapısı veya puanlayıcılara verilen eğitim bunlar arasında gösterilebilir (Weigle, 2002).

Ölçme süreçlerinde kullanılan araçlar ve yöntemleri farklı hata kaynakları etkileyebilir. Tüm bu kaynaklardan gelen hatalar ölçümlerden elde edilen sonuçların güvenirliğini etkiler. Ölçümlerden elde edilen sonuçların güvenirliği genellikle, bir ölçme aracından yüksek puan alan öğrencilerin, aynı beceri ve yetenekleri ölçmek için tasarlanmış diğer araçlardan da yüksek puan alma eğiliminde olma derecesi olarak tanımlanmaktadır (Dunbar, Koretz, & Hoover, 1991). Güvenirlik, aynı

alandaki görevler arasındaki korelasyonun bir fonksiyonudur. Tüm şartlar eşit varsayıldığında (örneğin görevlerin uzunluğu gibi), aynı alandaki görevler arasındaki korelasyon ne kadar yüksek olursa, birey için kabul edilebilir derecede güvenilir bir ölçüm elde etmek yönünden gereken test o kadar kısa olur ve bu nedenle test süreçlerinde gereken süre ve masraf da o kadar az olur (Klein, Stecher, Shavelson, McCaffrey, Ormseth, Bell, Comfort & Othman, 1998). Dolayısıyla ölçümlerin güvenirliği test hazırlama süreci için de ekonomik açıdan önemli bir yer tutmaktadır. Puanlayıcılar, yani ölçme aracına bireylerin verdiği yanıtlar doğrultusunda bir değer atayan uzmanlar, değerlendirmenin objektif olarak yapılabilmesi ve dolayısıyla güvenilir ölçümler için hata kaynaklarının önemli bir kısmını oluşturabilir. Çoktan seçmeli testler gibi yanıtı belli ve puanlaması nispeten pratik olan ölçme araçlarında dahi dikkatsizlik gibi nedenlerle puanlama hataları yapılabilir. Günümüzde teknoloji endüstrisinin gelişimi ve eğitime entegrasyonunda yaşanan hızlı gelişmeler bu tür kısa cevaplı veya çoktan seçmeli testlerde yapılabilecek hataları minimize etmiştir. Örneğin kompozisyon cevaplarının sanal ortamda puanlanması konusunda yapılan çalışmalar yapılandırılmış yanıtlı değerlendirmeleri uygun hale getirmiştir (Thorndike ve Thorndike-Christ, 2017). Amerika Birleşik Devletlerinde bu tür puanlama üzerine yapılan çalışmalar son zamanlarda çok revaçtadır ve bu tür puanlamaları yapabilen e-rater, c,rater, SpeechRater gibi paket programlar şirketler tarafından geliştirilmekte ve satışa sunulmaktadır (Atılgan, Aydın ve Kan, 2017). Bu tip puanlama yazılımları genel olarak noktalama, yazım ve imla hataları, gramer hataları, kullanılan kelime sayıları, organizasyon ve geliştirme aşamaları gibi etkenleri dikkate alarak hem kategorik hem de toplam puan oluşturma prensibiyle çalışmaktadır.

Performans ölçmelerinde kısa cevap veya çoktan seçmeli testler yerine değerlendirmesi nispeten daha zor olan araçlar kullanılır. Çünkü performans ölçümünde o performansa yönelik ölçme araçlarının seçilmesi gerekmektedir (Tekin, 1991). Dolayısıyla puanlamada yapılabilecek hataların olasılığı da artmış olur. Örneğin puanlama konusunda yeterli uzmanlığa sahip olmayan veya yeterince açık olmayan puanlama kriterleri kullanıldığında performans testlerinden alınacak ölçümlerin objektifliği ve güvenirliği, yanıtın seçildiği performans testlerindeki ölçümlere göre büyük oranda olumsuz etkilenecektir (Thorndike ve Thorndike-Christ, 2017 s. 323). Böyle durumlarda “Puanlayıcı Güvenirliği” kavramı karşımıza

çıkmaktadır. Bir puanlama aracıyla yapılan güvenilirlik kanıtı türü, puanlayıcılar arası güvenirliktir (Cook ve Beckman, 2006). Puanlayıcılar arası güvenilirlik, puanlayıcıların yaptıkları puanlama işleminde ne derece tutarlı olduklarını değerlendiren istatistiksel bir tekniktir. James, Demaree ve Wolf (1986) matematiksel olarak puanlayıcı güvenirliğini genellikle puanlamadaki toplam varyans ile puanlama setindeki sistematik varyansın oranı olarak tanımlamıştır. Puanlayıcı güvenirliği, iki bağımsız puanlayıcı tarafından (inter-rater) ve farklı zamanlarda aynı puanlayıcı tarafından (intra-rater) atanan puanların tutarlılığıdır (Reddy ve Andrade, 2010). Puanlayıcılar arası güvenirlik genellikle korelasyon veya varyans indislerinin analizi ile ifade edilmektedir (Fleenor, Fleenor ve Grossnickle, 1996). Yazma becerilerinin ölçülmesinde güvenirlik, genellikle iki veya daha fazla puanlayıcının bir yazma görevini puanlaması ve verdikleri puanlarda ne ölçüde uyum sağladıklarını gösteren puanlayıcılar arası uyum ile birbirinin yerine kullanılmaktadır. Bunun yanında puanlayıcı-içi uyum ise, tek bir puanlayıcının zaman içinde ne kadar tutarlı olduğuna yönelik bir göstergedir (Huebner ve Skar, 2021).

Puanlayıcılar arası güvenirlik katsayısı, puanlayıcılar arasındaki anlaşmanın/uyumun miktarını belirleyen istatistiksel bir ölçüdür (Gwet, 2011). Puanlayıcı güvenilirliğinin yüksek olması demek aslında ilgili puanlayıcıların birbiriyle veya farklı test ortamlarında güvenle değiştirilebilmesi ve araştırmacıların puanlayıcıdan gelecek hata kaynakları konusunda rahat olmaları anlamına gelmektedir. Bu kavram “birbiriyle değiştirilebilirlik” olarak ortaya çıkar (Gwet, 2014). Puanlayıcı güvenirliği için yapılan kestirimler, puanlıyıclar tarafından nispeten tutarlı bir şekilde puan verilip verilmediğini ele almak için kullanılır. Burada önemli bir husus, verilen puanların birebir eşit olması gerekmediğidir (Lebreton ve Jenell, 2008). Literatürde “puanlayıcılar arası uyum” (Inter-rater agreement) olarak karşılaşılan bir diğer kavram ise puanlayıcıların tam uyumuna işaret eder. Buradaki uyum tüm puanlayıcıların aynı bireyin aynı maddesine neredeyse aynı puanları vermesi iken puanlayıcı güvenirliği kavramı puanlayıcıların birbiriyle olan tutarlılığını ifade eder (Kozlowski ve Hattrup, 1992; LeBreton ve Jenell, 2008; Bliese, 2000). Bu durumdan çıkarılabilecek bir sonuç; puanlayıcılar arası güvenirliğin yüksek olması puanlayıcıların birebir uyum içinde olduğu sonucunu göstermemektedir. Araştırmacıların uyum ve güvenirlik arasından çalışmalarına uygun olanı seçmeleri

yerinde olacaktır. Tinsley ve Weiss (1975, s367) puanlayıcılar arası uyum için daha önce yapılan çalışmaları geliştirerek bir *"T"* puanı oluşturmuştur. *"T"* puanı 0 ve 1 arasında değerler almaktadır ve 0 değerini aldığında gözlenen şans uyumunun beklenen şans uyumuna eşit olması ve 1 değerini aldığında harika puanlayıcılar arası uyum yakalanması anlamına gelmektedir. İlgili eşitlik aşağıda gösterilmiştir: (Tinsley ve Weiss, 1975, s367)

$$\boldsymbol{T = \frac{N_1 - NP}{N - NP}} \quad (1)$$

$\mathbf{N_1}$ = Uyum sağlayan puanların sayısı

$\mathbf{N}$ = Puanlanan birey sayısı

P = Birey üzerinde şans uyumu olasılığı

Öğretmenler sınıflarda kullandığı ölçme araçlarını genellikle kendileri puanlarlar. Ancak yapılan araştırmalar birden fazla puanlayıcıdan yararlanmanın manidar derecede daha avantajlı olduğunu göstermektedir. Thorndike ve Thorndike-Christ (2017) bu avantajları şöyle sıralamıştır: Öncelikle puanlama işlemi sonunda puanlayıcılar arasında güvenirlik yorumları yapabilmek için karşılaştırma yapılabilir. Verilen puanlar birbirleri ile tutarlı ise puanlayıcıların güvenilir olduğu sonucuna ulaşılabilir. Puanlayıcıların uyumsuz olması durumunda puanlama yönergeleri ve anahtarları ile puanlayıcıların uzmanlıkları hakkında bir terslikten bahsedilebileceği gibi test maddeleri hakkında da bir araştırma yapılarak sorunun nereden kaynaklandığına dair bulgular ortaya konabilir. Aynı araştırmacılara göre bir diğer avantaj ise birden fazla puanlayıcının verdiği puanlardan ortak bir sonuca ulaşılabilir ve bu sayede bir puanlayıcının yapacağı hatayı diğerleri telafi etme şansına sahip olabilecektir. Görüldüğü gibi tek puanlayıcıya dayanan sonuçlarda ilgili değişkenler hakkında yorum yapmak veya bir yargıya varmak nispeten daha zordur. Bu araştırmada iki puanlayıcı yöntemi kullanıldığı için tek puanlayıcıya dayalı olan iç tutarlılık üzerinde durulmayacaktır.

Bir ölçme işleminde hangi tür aracın kullanıldığı ölçülmek istenilen özelliğin ne derece hatalardan arınık olduğuyla doğrudan bir ilişki içerisindedir. Bununla birlikte, Yazma görevleri, açık uçlu maddeler, performans görevleri gibi yanıtın birey tarafından oluşturulması istenen bazı ölçme araçlarında puanlayıcıdan kaynaklı çeşitli hataların sürece dâhil olması daha olası bir durumdur. Çünkü ne kadar iyi

ölçütler belirlenirse belirlensin yazma becerisini ölçen araçların puanlanması bir miktar öznel olmaktadır. Bu durum bu tip araçların puanlayıcı güvenirliklerini olumsuz etkiler (Tekin, 1991). Nitko ve Brookhart (2016, s.74) bu tip araçların puanlanmasında şu üç sorunun akla geleceğini belirtmiştir: 1. Farklı puanlayıcılar aynı yanıtları puanlarsa puanlar değişir miydi? 2. Sınav ne ölçüde objektif? 3. Farklı puanlayıcılardan elde edilen puanlar birbirlerinin yerine kullanılabilir mi? Bu durum için bir çözüm önerisi, puanlama yapacak iki ya da daha fazla kişinin bu konuda eğitim almış olması ve bu puanlayıcıların bağımsız olarak verdiği puanların ortalamasını bireyin son puanı olarak atamak olabilir (Atılgan, Aydın, Kan, 2017). Yazılı görevlerin puanlanmasında bazı puanlayıcı etkileri ile karşılaşmak mümkündür. Farklı puanlayıcı etkileri şu şekilde özetlenebilir (Knoch, Read ve Randow, 2007):

- *Katılık etkisi,* puanlayıcıların diğer puanlayıcılara veya önceden belirlenmiş ölçütlere kıyasla tutarlı bir şekilde ya çok katı ya da çok hoşgörülü puanlamada bulunmasıdır bulunmasıdır.
- *Hale etkisi, p*uanlayıcıların kavramsal olarak farklı kategoriler arasında ayrım yapamadığı, bunun yerine bir adayın performansını genel izlenime göre puanladığı ve böylece bir dizi farklı puanlama ölçeğinde aynı puanı verdiklerinde ortaya çıkan etkidir.
- Merkezi eğilim etkisi, Landy ve Farr (1983) tarafından "aşırı (olumlu veya olumsuz) puanlamalardan kaçınma veya ölçeğin orta noktasında veya yakınında puanlama yapma baskınlığı" olarak tanımlanmıştır.
- Tutarsızlık etkisi, bir puanlayıcının bir veya daha fazla puanlama ölçeği kategorisini diğer puanlayıcıların aynı ölçeği uygulama biçimine göre tutarsız bir şekilde uygulama eğilimi olarak tanımlanır.
- Yanlılık etkisi, puanlayıcıların puanlama durumunun bir yönüne göre alışılmadık şekilde katı veya hoşgörülü bir şekilde puanlama eğiliminde olduklarında ortaya çıkan etkidir.

İlgili araçlarda puanlama yaygın olarak; genel izlenim, sınıflama, sıralama ve güvenirliği en yüksek kabul edilen puanlama anahtarları kullanılarak yapılmaktadır (Baykul ve Turgut, 2009). Diğer üç yöntemde bağıl bir değerlendirme söz konusu iken anahtarla puanlama yönteminde bireyin verdiği yanıtlar diğer bireylerden

bağımsız olarak daha objektif bir biçimde değerlendirilebilir. DPA (bir diğer adıyla rubrikler) günümüzdeki anlamıyla, bireyin oluşturduğu yanıtların niteliğini değerlendirmek için kullanılan puanlama rehberidir ve yazılı kompozisyonlar, sözlü sunumlar ile bilim projeleri gibi görevlerde kullanılmaktadır. DPA'lar değerlendirme kriteri, nitelik tanımları ve puanlama stratejisi olmak üzere üç önemli özelliğe sahiptir (Popham, 1997).

Yazma görevi gibi performans görevlerinin değerlendirilmesinde kullanılan DPA, örtük veya açık olarak testin dayandığı teorik temeli temsil etmektedir. Bu, test geliştiricisinin test tarafından hangi becerilerin veya yeteneklerin ölçüldüğüne dair fikrini somutlaştırması anlamına gelir. Bu nedenle, bir DPA'nın geliştirilmesi ve her ölçek düzeyi için göstergeler belirlenmesi, ölçme sürecinin geçerliği bakımından kritik öneme sahiptir (Weigle, 2002).

Tüm bu süreçler puanlamanın dolayısıyla gözlenen özelliğin geçerlik ve güvenirliğini olumsuz etkileyebilir.

Puanlayıcı güvenirliği kestiriminde kullanılan bazı modeller. Tüm bu süreçler puanlamanın dolayısıyla gözlenen özelliğin geçerlik ve güvenirliğini olumsuz etkileyebilir. Puanlayıcılar arası güvenirlik, görevleri veya maddeleri doğru bir şekilde puanlamakla görevlendirilen puanlayıcıların tutarlılığını belirlemeye yönelik bir tekniktir. Bu tür güvenirliği değerlendirmenin bir dizi istatistiksel yolu vardır. Örneğin, puanlayıcıların birbirleriyle tam olarak kaç kez aynı puanı verdiklerinin yüzdesi olan uyum yüzdesi hesaplanabilir (Graham ve Milanowski ve Miller, 2012). Yaygın olarak kullanılan yöntemlerden bazıları arasında kesin uyum yüzdesi - Cohen'in kappası (1968) ve bu modelden türetilmiş bazı varyasyonları mevcuttur. Bunun yanında bir de sınıf içi korelasyon katsayısı bulunur (Cook ve Beckman, 2006). Bu yöntemler yaygın olmakla birlikte, Gwet'in(2002) AC1-2 istatistiği gibi ek yöntemler de puanlayıcılar arası güvenirliği belirlemek için bir araç olarak kullanılabilir. Diğer bazı yöntemler; sınıf içi korelasyon katsayıları (Intra-class correlation), genellenebilirlik kuramı ve Phi katsayısıdır. Bu istatistiklerin büyük bir bölümü duruma bağlıdır ve puanlayıcı sayısı gibi bazı faktörlerle belirlenir (Graham ve diğerleri, 2012). Bu istatistiklerin çoğu $\frac{(\boldsymbol{p_a}-\boldsymbol{p_e})}{(\mathbf{1}-\boldsymbol{p_e})}$ formundadır. Burada $\boldsymbol{p_a}$ uyumun yüzdesini ifade ederken $\boldsymbol{p_e}$ şans uyumunun yüzdesini ifade eder. Bu katsayılar genellikle uyum yüzdesinin genel bir formülleştirmesine dayanır ve bu formüller şans

uyumu yüzdesi bakımından birbirinden önemli ölçüde farklılaşır (Gwet, 2011). Bu yöntemlerden araştırmanın problemiyle ilişkili olanlar bu bölümde sunulacaktır.

Sosyal bilimlerde ve eğitim bilimlerinde ölçme işlemleri doğrudan yapılamaz. Sözgelimi, eğitim süreci başında bireyde bulunan özellikleri belirlemek veya bir öğretim süreci sonunda bireyin ne derece hedeflere ulaştığını kestirmek için ölçme araçları kullanılır. Bu araçlara; ortmandan, bireyden, ölçmeyi gerçekleştirenlerden veya kaynağı bilinmeyen birçok hata karışma olasılığı vardır. İdeal durum hatasız ölçme işlemi yapmak olsa da bu pratikte pek mümkün olmamaktadır. Dolayısıyla olabilecek en iyi çözüm bu hataların kaynağını bulmak ve en aza indirmektir. Genel olarak bu kavramdan yola çıkılarak oluşturulan Klasik test kuramında (KTK) (Lord ve Novick, 1968), psikometrik araçlarla gözlenen puanların (X), ölçme hatası (E) olmasaydı elde edilecek kişinin gerçek puanını (T) temsil edeceği düşünülmektedir. Bu durum eşitlik olarak ifade edilecek olursa:

Gözlenen Puan=Gerçek Puan + Ölçme Hatası

ya da gösterim olarak

$$X=T+E \tag{2}$$

gibi bir gösterimde bulunulabilir. Klasik test kuramında güvenirlik kestirimleri yapılırken varyans(Var) değerleri kullanılır. O halde eşitliği;

$$Var(X)= Var(T)+Var(E) \tag{3}$$

yani gözlenen puan varyansını, gerçek puan ve hata puanlarının ilişkisiz olduğu varsayımı sağlandığında, gerçek puan varyansı ile hata puanı varyansının toplamı olarak da ifade edebiliriz. Yukarıda da bahsedildiği gibi ölçme hatası *(E)* bireyin gerçek puanının *(T)* gözlenmesine engel olur ve bu hata kaynakları ölçme durumundan, puanlayıcıdan, bireyden, ortamdan, ölçme aracından veya kaynağı belli olmayan durumlardan oluşabilir. Tüm bu durumlar güvenirliği olumsuz etkileyebilir.

Puanlayıcı güvenirliğini belirlemek için kullanılan analiz yöntemleri puanlayıcılar arasındaki ölçme hatasından kaynaklanan varyans kaldırıldıktan sonra gözlenen puanlardaki varyansın ne kadarının gerçek puanlardaki varyansa bağlı olduğunu belirlemeyi amaçlamaktadır (Novick, 1966). Bu ifade:

$$\text{Güvenirlik} = \frac{Var(T)}{Var(X)} = \frac{Var(X) - Var(E)}{Var(X)} = \frac{Var(T)}{Var(T) + Var(E)} \quad (4)$$

Bir örnekle açıklamak gerekirse, 0.90'lık bir puanlayıcı güvenirliği sonucu gözlenen varyansın %90'ının gerçek puan varyansına veya puanlayıcılar arası benzerliğe işaret ettiğini ve kalan %10'luk kısmında hata varyansı veya puanlayıcılar arası farklılığa işaret ettiği yorumu yapılabilir.

Puanlayıcı güvenirliği hesaplamalarında dikkat edilmesi gereken bir diğer nokta araştırma yapılacak veri setine uygun modelin seçilmesidir. Bu model seçilirken dikkat edilecek noktalardan bazıları verilerin ölçeği(sınıflama, sıralama, eşit aralık ve eşit oran), puanlayıcı sayısı ve puanlama desenidir. Örneğin aynı yanıtı farklı puanlayıcılar puanlayabilir ya da farklı bireylerin yanıtlarına aynı puanlayıcılar değer atayabilir. Bu araştırmada farklı bireylerin yanıtlarını iki puanlayıcı yanıtlayacağı için bu desene yönelik modellerden bazılarına kısaca değinilmiştir.

Puanlayıcı güvenirliğine dair istatistiklerden birisi Cohen'in ağırlıklandırılmış Kappa katsayısıdır (Cohen, 1968). Jacob Cohen'in puanlayıcılar arası güvenirlik için 1960 yılında yaptığı çalışma sadece sınıflama ölçeği düzeyindeki veriler için uygundur. Şansa bağlı uyum etkisi çıkarılarak nihai bir uyum puanı ortaya çıkar. Örneğin 1,2 ve 3 puanlarından oluşan kategorilerle puanlanan verilerde birinci puanlayıcı öğrencinin yanıtına 3, ikinci puanlayıcı ise 2 puan vermiş olabilir. Burada aynı puan verilmediği için kesin bir uyumsuzluk söz konusudur. Ancak aynı araştırmacının 1968 yılında yaptığı çalışmada puanlayıcıların aynı puanları verip vermediklerinden ziyade puanların düzeyleri (kötü, orta ve iyi) de istatistiksel modele girmiştir (Hallgreen, 2012).

Puanlayıcıların ne oranda uyum sağladıklarının yüzde olarak ifadesi geçmişte çokça kullanılmış olsa da bu yöntemin bir sorunu, uyumun şansa bağlı olarak oluşabileceğidir (Cohen, 1968). Yani böyle bir uyum esasen tesadüfi olarak ortaya çıkan uyumu hesaba katmamaktadır. Cohen'in Kappa'sı şans eseri oluşan uyumun dışında kalan gerçek uyumu göstermektedir. Bu durum şöyle ifade edilmektedir:

$$K = \frac{gözlenen\ uyum - tesadüfi(şans)\ uyumu}{1 - tesadüfi(şans) uyumu} \quad (5)$$

Aynı eşitlik P_0 gözlenen uyum ve P_e beklenen şans uyumu olmak üzere

$$K = \frac{P_0 - P_e}{1 - P} \quad (6)$$

Özellikle performans ölçme yaklaşımlarında kesin bir doğru ya da yanlış yanıttan ziyade öğrencinin ürünü oluştururken hangi kriterlere dikkat ettiğini belirlemek ve buna uygun bir toplam puan oluşturmak daha uygundur (Atılgan, Kan, Doğan, 2009). Örneğin bir paragraf yazma görevi en düşük 1 ve en yüksek 3 olmak üzere 2 puanlayıcı tarafından puanlansınmak üzere dört alt kategori olsun. Değerlendiriciler bu kategorilere iyi (3 puan) orta (2 puan) ve yetersiz (1 puan) olmak üzere değerler verdiğinde oluşacak örnek yapı tablo 1'de gösterilmiştir:

Tablo 1

Örnek Paragraf Görevi Puanlama Tablosu

	M1	*M2*	*M3*	*M4*	*M1*	*M2*	*M3*	*M4*
A	1	2	2	3	1	3	2	1
B	2	2	3	3	2	2	2	2
C	3	3	1	2	2	1	3	3
D	2	1	2	1	3	3	1	1
E	1	1	2	1	1	2	2	3
F	3	3	1	2	2	1	2	2

Tablo 1'de görüleceği gibi, değerlendirme kategorileri "iyi-3", "orta-2" ve "yetersiz-1" olarak verilmiş puanlardır. Bu kategoriler sıralama ölçeğindedir ve paragraf görevinden alınan puanları göstermektedir. Burada ürünün iyi veya yetersiz olarak puanlanmasının dışında bir de orta seviyede derece mevcut olduğu için iki puanlayıcının uyumsuzluğundan ziyade bu uyumsuzluğun derecesi önemlidir. Bu uyumsuzluk derecesini ağırlıklandırmak için yeni bir ağırlıklandırılmış Kappa katsayısı geliştirilmiştir. Birçok ağırlıklandırma yöntemi bulunmakla birlikte yaygın olarak ikinci derece ağırlıklandırma kullanılmaktadır (Sim ve Wright, 2005). Ağırlıklandırılmış κ katsayısının hesaplanması için:

$$\kappa_{\omega} = 1 - \frac{\sum_{i=1}^{\kappa} \sum_{j=1}^{\kappa} \omega_{ij} x_{ij}}{\sum_{i=1}^{\kappa} \sum_{j=1}^{\kappa} \omega_{ij} \varepsilon_{ij}} \quad (7)$$

burada *k* kod sayısına, ω_{ij} , x_{ij} ve ε_{ij} sırasıyla ağırlık matrisi, gözlenen matrisi ve beklenen matrislerindeki faktörlere (i-inci satır ve j-inci sütun) karşılık gelir (Reis ve Judd, 2014).

Kappa katsayısının bir sınırlılığı, puanlayıcılar arası uyumsuzluğun şansa mı yoksa sistematik farklılıklara mı bağlı olduğu hakkında bir bilgi vermediği ve yorumun da buna göre yapılması gerektiğidir (Hartman, 1977). Mükemmel uyum (κ = 1), puanlayıcılar arasında anlaşmazlık olmadığında, yani tam uyum sağladığında elde edilir. Sıfır değeri, iki ortalama arasında gözlenen ortalama uzaklığın yalnızca şans eseri uyuma bağlı olduğunu gösterirken, negatif değerler ise, gözlenen ortalama uzaklığın şans eseri beklenen ortalama uzaklıktan daha büyük olduğunu ifade eder.

κ tipi katsayıların yorumlanmasında genel olarak Tablo 2'de bulunan uyum aralıkları kullanılabilir (Cicchetti, 1994):

Tablo 2

Kappa Katsayısının Kullanılmasında Uyum Aralıkları

Kappa (κ)	Uyumun Gücü
< 0,40	Zayıf
0,40 – 0,59	Yeterli
0,60 – 0,74	İyi
0,75 – 1,00	Mükemmel

Puanlayıcılar arası güvenirlik kestiriminde bir diğer yöntem Krippendorf'un Alpha ($\boldsymbol{a}$) katsayısıdır (Krippendorff, 1980). ($\boldsymbol{a}$), puanlayıcılar arasındaki uyumu ölçmek veya bunlara hesaplanabilir değerler atamak için geliştirilmiş bir güvenilirlik katsayısıdır. Bu katsayının genel formu şu şekilde ifade edilir (Krippendorff, 1980, 2011):

$$a = 1 - \frac{D_o}{D_e} \quad (8)$$

Burada $\boldsymbol{D_o}$ gözlenen uyumsuzluğu ifade ederken $\boldsymbol{D_e}$ ise gözlenen uyumsuzluğun şansa uyumuna atfedilen kısmını belirtir. Matematiksel olarak düşünüldüğünde $\boldsymbol{D_o}$=0 olduğunda $\boldsymbol{\alpha}$=1 değerini alır yani bu mükemmel uyum ve güvenirlik anlamına gelir. Aynı şekilde puanlayıcılar şansa bağlı olarak uyum gösterdiklerinde $\boldsymbol{D_o}$=$\boldsymbol{D_e}$ olur ve $\boldsymbol{\alpha}$=0 değerini alır ve bu durum güvenirliğin yani uyumun yokluğuna işaret eder. $\boldsymbol{\alpha}$=0 durumu puanlayıcıların birimler arasında ayrım yapamadıklarında veya örneklemin toplu bir kestiriminden rastgele alınan değerleri onlara atayamadıklarında ortaya çıkar. Herhangi bir yöntemle üretilen verilere güvenmek için, $\boldsymbol{\alpha}$ 'nin bu iki uç durumdan uzak olması gerekir ve ideal olarak $\boldsymbol{\alpha}$=1 olmalıdır (Krippendorff, 2011).

Krippendorff'un a katsayısı bilinen diğer güvenirlik katsayılarının aksine her türlü puanlayıcı sayısına, sınıflama, sıralama, aralık veya oran türünde her türlü ölçek türüne, her sayıda kategoriye, eksik veya kayıp veri bulunan durumlara ve minimum bir sayı gerektirmeksizin büyük veya küçük tüm örneklem gruplarına uygulanabilir olmasıyla işlevseldir (Krippendorff, 2004; s222).

Krippendorff'un a katsayısı değeri hesaplanırken bütün veri tipleri için izlenen yol benzerdir. Veriler uyum (coincidence) matrisi olarak adlandırılan simetrik bir aktarılır. Bu matrisin oluşturulma şekli ölçek tiplerine göre farklılık göstermektedir. Uyum matrisinin oluşturulmasının ardından metrik farklara dayalı olarak gözlenen ve beklenen uyumsuzluk hesaplanır ve bu değer 1'den çıkarılarak Krippendorff'un a değeri bulunur.

Sıralama ölçeğinde iki puanlayıcı ile yapılacak güvenirlik kestirimine ait detaylı hesaplama şu şekilde yapılmaktadır:

Eşitlik (8) hatırlanacak olursa:

$$a = 1 - \frac{D_O}{D_e}$$

D_O Gözlenen uyumsuzluk ve D_e beklenen uyumsuzluk ve δ^2 uygun ölçeğin bir fonksiyonu olan herhangi iki c ve k değeri arasındaki kare farkı, O_{ck} gözlenen c ve k değerleri olmak üzere,

$$D_o = \frac{1}{n}\sum_c \sum_k O_{ck\,\text{ölçek}}\delta_{ck}^2 \tag{9}$$

$$D_e = \frac{1}{n(n-1)}\sum_c \sum_k O_{ck\,\text{ölçek}}\delta_{ck}^2 \tag{10}$$

Tüm *N* maddesinin en az iki puanlayıcı tarafından puanlandığı varsayılsın. Bütün ölçek türlerinde ve iki veya daha fazla puanlayıcı ve kategori olduğu durumda (Zapf, Castell, Morawietz, ve Karch, 2016):

$$\mathrm{A} = 1 - \frac{D_o}{D_e} = \frac{\sum_j \sum_{j'} o_{jj'}\delta_{jj'}^2}{\sum_j \sum_{j'} e_{jj'}\delta_{jj'}^2} \quad (\text{kategori } j, j' = 1 \ldots, k), \tag{11}$$

D_O Gözlenen uyumsuzluk ve D_e beklenen uyumsuzluk olmak üzere, gözlenen uyum şu şekilde ifade edilir:

$$o_{jj'} = \sum_{i=1}^{N} \frac{i \text{ maddesindeki } j - j' \text{ eşleşmelerinin sayısı}}{n_i - 1} \tag{12}$$

n_i, *i* maddesini puanlayan puanlayıcıların sayısını ifade eder. Buna karşılık, şansa bağlı beklenen uyum $j, j' = 1, \ldots, k$. kategorisindeki puanlama sayısı $n_j, n_{j'}$ olmak üzere şu şekilde ifade edilir:

$$e_{jj'} = \begin{cases} n_j\left(n_{j'} - 1\right)/\left(\sum_i n_i - 1\right) & \text{if } j = j' \\ n_j n_{j'}/\left(\sum_i n_i - 1\right) & \text{if } j \neq j' \end{cases} \tag{13}$$

$j, j' = 1, \ldots, k$ kategorileri için ölçeğe özgü fark fonksiyonu olan $\delta_{jj'}^2$ fonksiyonu ölçeğin türüne göre aşağıdaki gibi ifade edilmektedir:

Sınıflama düzeyinde ölçek: $\delta^2_{jj'} = \begin{cases} 0 \text{ if } j = j' \\ 1 \text{ if } j \neq j' \end{cases}$

Sıralama düzeyinde ölçek: $\delta^2_{jj'} = \left(\frac{n_j}{2} + \sum_{g>j}^{g<j'} n_g + \frac{n_{j'}}{2}\right)^2 , \ j < j'$

Eşit aralık düzeyinde ölçek: $\delta^2_{jj'} = (j - j')^2$

Eşit oran düzeyinde ölçek: $\delta^2_{jj'} = \left(\frac{j-j'}{j+j'}\right)^2$

Krippendorff α katsayısının istenilen değer aralıkları ve yorumu(Krippendorff, 1995) Tablo 3'te gösterilmiştir (Akt. Bilgen ve Doğan, 2017):

Tablo 3

Krippedorff α Katsayısının Değer Aralıkları ve Yorumu

α	Uyumun Gücü
< 0,67	Zayıf
0,67 – 0,80	Orta
0,80 ≤	Yüksek

Genellenebilirlik (G) kuramı. Güvenirlik ilk defa Spearman (1904) tarafından bir korelasyon ölçüsü olarak ortaya atılmıştır. Güvenirlik kavramı genel olarak hatırlanacak olursa, bir ölçme işleminin sonucunda kararın dayandırılacağı puan, aynı amaca hizmet edebilecek birçok ölçümden yalnızca birisidir. Kararın dayandırılacağı ideal veri, kişinin tüm kabul edilebilir gözlemler üzerindeki ortalama puanına benzer bir puan olacaktır (Cronbach, Gleser, Nanda ve Rajaratnam, 1972). Yani güvenirlik kavramı aslında bireyin gözlenen puanının, aynı bireyin eşit şartlarda aynı testi sınırsız olarak aldığında oluşacak puan ortalamasına genellenebilmesi olarak da ifade edilebilir (Shavelson ve Webb, 1991). Genel olarak düşünüldüğünde, güvenirlik, gözlenen puanlardaki tutarlılık ve tutarsızlıkların derecesini içerir (Brennan, 2001). Fakat bu sonsuz test uygulaması ve bunların ortalamasının alınması durumu pratikte mümkün olmamaktadır. Araştırmacılar bu noktada bireye yapılacak test sonucunun olası tüm sonuçlara yani evrene genellenebilmesi için o testi en iyi şekilde tasarlama çalışmaları yürütmüşlerdir.

Özellikle sosyal bilimlerde ölçme ve değerlendirme yaparken doğrudan gözlem imkânı bulunmadığı için bazı araçlar kullanılmaktadır ve bu sürece ister

istemez farklı kaynaklardan hatalar karışması kaçınılmazdır. KTK'da güvenirliğin hesaplanması anlamında kullanılan yöntemlerden test-tekrar test yöntemine göre hatanın kaynağı zaman, eşdeğer formlar yöntemine göre hatanın kaynağı formlar ve son olarak iç-tutarlılığı belirleyen yöntemlere göre de testin maddeleri hata kaynağı olarak göstermektedir (Atılgan, 2019). KTK'da güvenirliğe dair hesaplamaların amaca hizmet edeceği şekilde uygun bir yol seçilerek hesaplanması gerekmektedir (Baykul, 2000). KTK'nın belki de en büyük sınırlılıklarından birisi hata kaynaklarını bir bütün olarak ele almasıdır. Çünkü KTK, güvenirlik kestirimleri yaparken tek bir hata oranı belirler ve bu hatanın, örneğin bir puanlama işleminde, puanlayıcıdan mı, bireyden mi yoksa araçtan mı kaynaklandığı konusunda bir bilgi vermez (Shavelson ve Webb, 1991). Klasik anlamda güvenirlik, bir seferde yalnızca bir tür hatanın kestirimine (örneğin puanlayıcılar, görevler veya durumlar arası olmak yerine yalnızca formlar arasındaki tutarsızlığa atıfta bulunma gibi) izin vermektedir. (Webb, Rowley ve Shavelson, 1988).

KTK'nın davranışların ölçülmesindeki bu sınırlı uygunluğu nedeniyle, bu tür hata kaynaklarını ele alabilen bir model olarak ve davranış değerlendirmesinin teknik bir yeterliğini oluşturmak amacıyla genellenebilirlik (G) kuramı geliştirilmiştir (Cronbach, Rajaratnam ve Gleser, 1963; Cronbach, Gleser, Nanda ve Rajaratnam, 1972; Cone, 1977). G kuramı, çok sayıda ölçme durumunu ele almak için kapsamlı bir kavramsal çerçeve ve güçlü bir dizi istatistiksel prosedür sunar ve bir dereceye kadar belirli varyans analizi (ANOVA) prosedürlerinin bu ölçme durumlarına uygulanması yoluyla klasik test teorisinin bir uzantısı olarak görülebilir (Brennan, 2001). Ölçmenin amacı, bir ölçümden genel sonuçlar çıkarmaktır (Sireci, 1998). G kuramı, bu görevi başarmak için, verilerin ölçme prosedürünün tüm önemli koşullarında genelleştirmeyi mümkün kılan bir yöntem kullanılarak toplanması gerektiğini öne sürmektedir (Wright ve Piersel, 1992). G kuramının kullanım amacı, belirli bir puanlayıcı tarafından elde edilen puanın, tüm olası test durumlarında ve tüm olası puanlayıcılarda tipik olarak kabul edilebilecek bir duruma ne kadar genellenebilir olduğunun belirlenmesidir. G kuramının gücü, bir ölçmedeki çoklu hata kaynaklarının ayrı ayrı kestirilebilmesine imkân sağlar. Ayrıca karar vericilere; güvenilir ölçümler elde etmek için ne kadar farklı durum, test formu ve puanlayıcıya ihtiyaç olduğunun göstergelerini de verir (Shavelson ve Webb, 1991). KTK ve ANOVA, G kuramının temelleri olarak sayılmasına rağmen, kuram bu iki temel

yapının basit birleşiminden hem daha fazla hem de daha azdır. Örneğin, KTK'nın ve ANOVA'nın tüm yönleri G kuramına dâhil edilmemiştir. Hatta ANOVA'nın bazı perspektifleri G kuramı ile tutarsızlık bile göstermektedir. G kuramı daha çok varyans bileşenleri ve bunların tahminine odaklanır (Brennan, 2001). Bu kuram ile birlikte test geliştiriciler ölçümlerin güvenirliğini çeşitli yorumlamalar yaparak araştırabilir. Örneğin KTK sonucuna dayalı bir yorum olan “Ali, grubun %50’sinden daha yüksek puan aldı.” İfadesinde test geliştiriciler Ali’nin içinde bulunduğu gruba yönelik bir çıkarım yapabilirler. Ancak G kuramı bireyin gruptan bağımsız olarak mutlak performansı üzerine bir ölçüm güvenirliği bilgisi sağlayabilir (Shavelson ve Webb, 1991). Performans değerlendirme ölçümlerinde G kuramının yararlı olduğu düşünülmektedir (Mitchell, 1979).

Genel olarak G kuramı KTK’yı şu önemli yönlerde genişletmektedir (Shavelson ve webb):

G kuramı öncelikle istatistiksel olarak tek bir analizde her bir hata kaynağının büyüklüğü hakkında tahminde bulunur ve ölçüm güvenirliğini en uygun hale getirmek için bir yöntem sağlar.

İkinci olarak, G kuramı "genellenebilirlik (G) katsayısı" olarak adlandırılan bir güvenirlik katsayısı sunsa da, kuram, ölçme işlemini etkileyen her bir hata kaynağının büyüklüğünü indeksleyen varyans bileşenlerine odaklanır.

Üçüncü olarak, G kuramı bağıl değerlendirme ile mutlak değerlendirme arasında bir ayrım yapar ve buna uygun katsayılar üretebilir. Bağıl değerlendirmeler, bir bireyin diğerinden "ne kadar daha iyi" performans gösterdiğiyle ilgilidir. Mutlak değerlendirmeye dair yapılan yorumlar, bir bireyin akranlarının performanslarına bakılmaksızın "ne kadar iyi" performans gösterebileceğine ilişkin değerlendirmeleri ele alır.

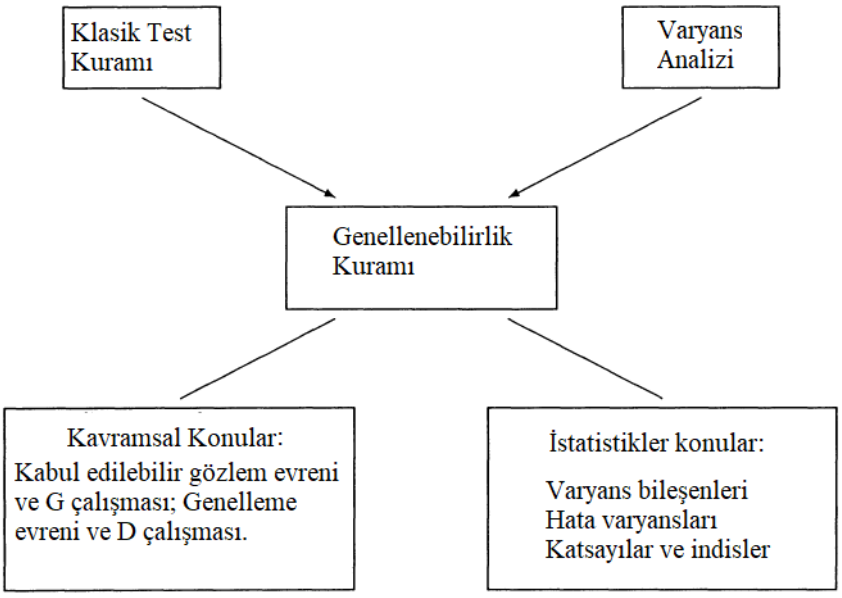
Dördüncü olarak, G kuramı, G ve D (karar) çalışmaları arasında ayrım yapar. G çalışmaları, mümkün olduğunca çok sayıda potansiyel ölçme hatası kaynağının büyüklüğünü tahmin etmektedir. D çalışmaları ise, uygulanan belirli bir amaç için hatayı en aza indiren bir ölçme durumu tasarlamak için G çalışmasından alınan bilgileri kullanır. G çalışmasında tahmin edilen varyans bileşenleri, daha sonra, bir D çalışmasında yaygın uygulama için zaman ve maliyet açısından verimli bir ölçme durumu tasarlamak için kullanılır.

G kuramı, bireyin gözlenen puanının bir evren puanından (KTK'daki gerçek puana karşılık gelen) ve bir veya daha fazla hata kaynağının bir araya gelmesiyle oluştuğunu varsayar. Ölçümler arasındaki değişkenlik; bireyler arasındaki farklılıklar ve diğer birçok hatadan kaynaklanabilir.

G kuramı, gözlenen puanlar için bir varyans bileşeni (σ^2) tanımlar. Genellikle ölçme nesnesi olan bireyler için varyans bileşeni (σ_p^2)'dir. Maddeler için varyans bileşeni (σ_i^2)'dir ve artık varyans bileşeni (residual) $\sigma_{pi,e}^2$'dir. Değişkenliği bu şekilde bölümlere ayırarak, G kuramı araştırmacının başlıca ölçme hatası kaynakları ve her bir hata kaynağının büyüklüğü için bir G katsayısı oluşturmasına olanak verir.

Bununla birlikte, Amerika Eğitimsel Araştırma Kuruluşu (AERA) gibi eğitim araştırmaları alanında standartlar belirleyen bazı büyük kurumlar, çeşitli noktalarda açıkça G kurumuna atıfta bulunur. Örneğin, tekrarlanan veya paralel ölçümlere dayalı güvenirlik tahminlerine ilişkin olarak standart 2.10'da; "Mümkün olduğunda, her kaynağın hata varyanslarının kestiriminde bulunulmalıdır. Genellenebilirlik çalışmaları ve varyans analizleri özellikle bu açıdan faydalıdır. Bu analizler, maddeler ve puanlayıcılar için ve kararlılık süresi içindeki durumlar için ayrı hata varyans kestirimlerinde bulunabilir." İfadesi buna örnek olarak gösterilebilir (AERA, 1999, s. 34).

G kuramı şekil 1'de de gösterildiği gibi, sadece hata kaynaklarıyla ilgili değil aynı zamanda verimli ölçme prosedürleri dizayn edilmesine de olanak verir.



Şekil 1. Genellenebilirlik teorisinin kuramsal çerçevesi ve temelleri

(Brennan, 2001)

Genellenebilirlik kuramı ile ilgili kavramlar. G kuramındaki bazı kavramları anlamak kuramın üzerine yerleştirildiği yapıyı bir bütün olarak görebilmek açısından önemlidir. Bu bölümde bu kavramlara yer verilmiştir.

Ölçme işleminin yapıldığı her ölçme durumunda bir veya birden fazla varyans, yani değişenlik kaynağı bulunmaktadır. G kuramında görev, puanlayıcılar ve zaman gibi benzer ölçme durum setlerine değişkenlik kaynağı adı verilir (Brennan 2001). Yapılan bir çalışma sonunda eğer evrene genelleme yapılacaksa değişkenliğin kaynağı tesadüfî, genelleme örneklem ile sınırlı kalacak ise bu durumda kaynak sabit olarak adlandırılır (Brennan, 2001). Bu değişkenlik kaynaklarının her birine Yüzey (Facet) denilmektedir (Shavelson ve webb, 1991). Brennan (2001); Crocker ve Algina (2006) yüzey'i basitçe bir dizi benzer ölçme koşulu olarak tanımlamaktadır. Örneğin, 10 maddelik bir bir yazma becerisi testinde her madde için performansın iyi, orta ve geliştirilmeli kategorilerine göre iki puanlayıcı tarafından puanlandığı varsayılsın. Burada ölçmenin yapıldığı iki koşul yani yüzey puanlayıcı ve performanstır. Yüzeyler aslında ANOVA'daki faktör kavramıyla aynı şeyi ifade ederler (Cardinet, Johnson ve Pini, 2009). Yüzey tanımı yapılırken göz önünde bulundurulması gereken bir husus, örnekten de anlaşılabileceği üzere, ölçmenin gerçekleştirildiği gruptaki bireylerin yüzey olarak adlandırılmamasıdır. Çünkü ölçmenin amacı bireylerdir. Dolayısıyla burada bireyler bir yüzey olarak ele alınmazlar. Bireylere atfedilen varyans istenen durum olduğundan aynı KTK'da hata kaynağı olarak alınmadığı gibi G kuramında da yüzey değildir. Burada bireylere "ölçme objesi" adı verilir. (Webb, Rowley ve Shavelson, 1989). Bununla birlikte test durumundaki yüzeylerin kendi içinde de düzeyleri mevcuttur. Bu düzeyler G kuramında "Koşul" (Condition) olarak adlandırılır. Örneğin iki puanlayıcının bulunduğu bir senaryoda puanlayıcı sayıları yani birinci puanlayıcı ve ikinci puanlayıcı koşul olarak adlandırılabilir. Koşullar da yine ANOVA'daki "level" yani düzeylere karşılık gelir (Shavelson ve Webb, 1991). Evren puanı, genelleme yapılacak evrendeki tüm koşullara ilişkin ölçme objelerine yönelik bir ortalama puandır. Tüm ölçme objeleri ise "evren puan varyansı" şeklinde adlandırılmaktadır. Bu, KTK'da bulunan gerçek puan varyansına benzer (Brennan, 1992, Shavelson ve Webb, 1991).

Kabul edilebilir gözlemler ereni ve genelleme evreni. G kuramının bakış açısından, bir ölçme kabul edilebilir gözlemler evreninin tek bir formudur. Burada ifade edilen gözlemler, birbirinin yerine geçebilen her gözlem olarak ele alınabilir (Shavelson ve webb, 1991). Örneğin, bir sınavda bir öğrencinin aldığı puan, performansının kabul edilebilir tek göstergesi değildir. Aynı testin farklı bir formundan veya aynı formun farklı bir maddesinden alınan bir puan gibi, farklı bir zamanda alınabilecek bir puan da kabul edilebilir (Alkharusi, 2012).

Genellenebilirlik çalışmaları, belirli koşullar altında ölçülerek de yapılabilir. Bu koşulların tipik olarak daha geniş bir koşul kümesini temsil ettiği kabul edilecektir. Tüm bu koşullarda alınabilecek gözlem popülasyonu, kabul edilebilir gözlemler evreni olarak adlandırılır (Crocker ve Algina, 2008). Aynı zamanda, Kabul edilebilir gözlemler evreni, ölçme uygulamasındaki değişkenlik kaynaklarının koşullarına karşılık gelen tüm olası koşulların kombinezonları demektir. Genellenmesi istenen bir yüzeye ait koşullar da "genelleme evreni" olarak adlandırılabilir (Webb, Rowley ve Shavelson, 1989). Yukarıdaki yazma görevi örneği hatırlanacak olursa, olası bütün iki puanlayıcı ve 10 maddeden oluşacak tüm durumlar genelleme evreni ve benzer durumlardaki olası tüm puanlayıcı ve maddelerin içerdiği evren de kabul edilebilir gözlemler evreni olarak tanımlanabilir. 10 Maddeden (ki bunlar madde yüzeyini oluşturur) ölçme işlemiyle belirlenmesi istenen beceriyi ölçebilecek madde evrenine genelleme yapılması istenir. Puanlayıcılar için de aynı durum geçerlidir. Puanlayıcı yüzeyi olan iki puanlayıcıdan yola çıkılarak tüm puanlayıcıları barındıran evrene genellenmesi beklenir. Daha önce de bahsedildiği gibi maddelerin ve puanlayıcılara ait yüzeylerin koşullarından genellenmesi beklenen evrenler genelleme evrenidir. Her araştırmacı farklı kabul edilebilir gözlemler evreni senaryoları oluşturabilir. Burada dikkat edilmesi gereken husus, yapılan çalışmaları yorumlarken ve değerlendirirken bunun göz önünde bulundurulması gerektiğidir. G kuramında evren kelimesi, ölçme koşullarını yani yüzeyleri ifade ederken popülasyon kelimesi ölçme objeleri için kullanılır (Brennan, 2001).

Genellenebilirlik (G) ve karar (D) çalışmaları G çalışması ve D çalışması, G kuramında farklı işlevleri yerine getirir, ancak iki çalışma, ölçmeyi geliştirme süreci için birbiri ile bağlantılı kalır. G çalışması önce gözlenen puanlardan farklı değişkenlik kaynaklarını sınıflandırarak tahsis eder. Olası hata kaynaklarının hepsi analize katılır ve bu kaynakların belirlenmesi, etkilerinin ortaya konması ve kabul

edilebilir gözlemler evreninin tanımlanması gibi işlemler yapılır (Webb, Rowley ve Shavelson, 1989). Tepkilerin zaman içindeki kararlılığı, bir aracın iki veya daha fazla formundaki puanların denkliği veya alt ölçek puanlarının veya bir ölçekteki maddelerin birbiriyle ilişkisi ile ilgili çalışmaların tümü G çalışmaları olarak kabul edilebilir (Crocker ve Algina, 2006). G çalışmasının temel amacı, belirli bir ölçme prosedürünün özelliklerini değerlendirmek ve ölçme kesinliğini tahmin etmektir. Bu amaçla ilk olarak farklı ölçme hatası kaynakları belirlenmelidir. Böylece hataya kaynaklarının derecesi ve göreceli önemi ölçülebilir. Bir G çalışması planlanırken, araştırmacı tarafından yapılması gereken ilk şey, ölçme sürecinde rol oynayan yüzeyleri ve bunların birbirleriyle olan ilişkilerini belirlemektir (Cardinet, Johnson ve Pini, 2009). Gözlenen puan ile evren puanı arasındaki ilişkiyi araştırmak için G çalışması gereklidir. Ardından yapılacak D çalışması ise, sosyal bilimler alanında belirli bir amaç için gerçekleştirilecek ölçme prosedürünün mümkün olan en iyi uygulamasını tasarlamak için G çalışması tarafından sağlanan bilgileri kullanarak bir karar verme amacı güder (Shavelson ve Webb, 1991). D çalışması genelleme evreninden yeni örnekler toplayarak hatayı gidermek için elde edilen varyans bileşenlerine dayanır ve ardından ihtiyaçları karşılayacak bir ölçme düzeni tasarlanmasına yardımcı olur. D çalışmasındaki temel amaç bir iyileştirme süreci oluşturmaktır. Bu adımda, G çalışmasında elde edilen sonuçlar, özellikle ölçme hatasına katkıda bulunan ana unsurlar için tahmini varyans bileşenlerini kullanılır. G çalışmasında olduğu gibi D çalışması da genellikle tek bir gözlemden ziyade ortalamaya odaklanır, bu nedenle D çalışmasının varyans bileşenleri, G çalışmasının varyans bileşenlerinin karşılık gelen yüzey (facet) düzeylerinin sayısına bölünmesiyle elde edilir. Hata yüzeyi seviyelerinin örneklem boyutu ne kadar büyükse, hata varyansı o kadar az olacaktır. Etkilerin basitleştirilmesi ve sınıflandırılması için G çalışmaları için genellikle çapraz desenler önerilir, ancak D çalışmaları gerektiğinde her türlü tasarımı kullanabilir (Shavelson ve Webb, 1991).

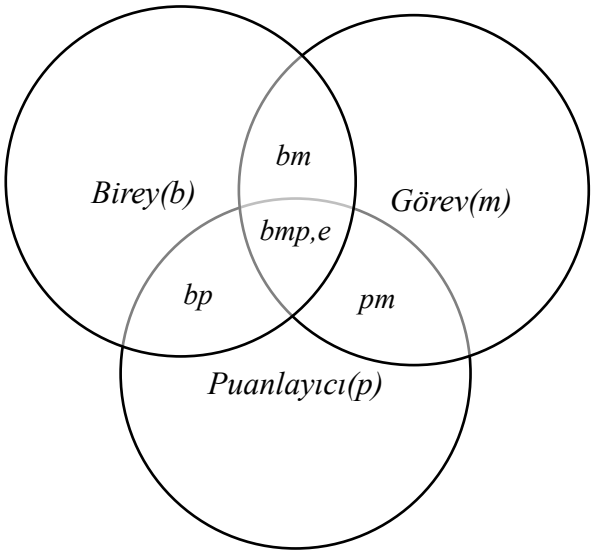
G kuramında kullanılan desenler. G kuramı çalışmalarında bulunan değişkenlik kaynaklarının sayısı ve ilgili desenin tasarımına yönelik faklı desenlerler oluşturulabilir. Bağlı olarak çaprazlanmış ya da yuvalanmış desenler ve karışık desenlerden oluşabilir. Bir araştırmada kullanılan değişkenlik kaynağına ait bütün yüzeyler diğer değişkenlik kaynağına ait yüzeylerin hepsinde bulunduğunda bu araştırma deseni *tümüyle çaprazlanmış (fully-crossed) desen* olarak anılır. Bu

desen *x* işaretinin yüzeylerin arasına konması şeklinde gösterilir (Shavelson ve Webb, 1991). Bir örnek ile açıklanacak olursa, geliştirilen bir yazma testinde testi alan tüm bireyler bireyler *(b)*, puanlayıcılar *(p)* ve testte yer alan yazma görevleri de *(m)* olsun. Testi alan 4 bireyin 1 ve 4 arasında puanlandığı 4 madde 2 farklı puanlayıcının tarafından puanlandığı varsayılsın. Bu puanlayıcılar testi alan tüm bireylerin cevaplarını puanladığında söz konusu desen çaprazlanmış desen olarak adlandırılırken bu desene çapraz desen denir ve *b x m x p* şeklinde gösterilir. Daha önce de bahsedildiği gibi eğitim bilimlerinde bireyler ölçmenin temel konusu yani *ölçme objesi* oldukları için değişkenlik kaynağı sadece puanlayıcılar ve maddelerdir. Söz konusu desen için bir örnek Tablo 4'te gösterilmiştir.

Tablo 4

Tümüyle Çaprazlanmış Desene İlişkin Örnek (bxmxpx)

	Madde1		Madde2		Madde3		Madde4	
Birey	Puanlayıcı 1	Puanlayıcı 2	Puanlayıcı 1	Puanlayıcı 2	Puanlayıcı 1	Puanlayıcı 2	Puanlayıcı 1	Puanlayıcı 2
1	2	3	4	4	2	2	3	3
2	2	3	1	4	3	3	1	3
3	1	2	3	1	2	3	3	2
4	1	1	4	3	2	2	3	1



Şekil 2. İki yüzeyli tümüyle çaprazlanmış desen modeli

Şekil 2'de iki yüzeyli tümüyle çaprazlanmış desene ait değişkenlik kaynakları modeli Venn şeması şeklinde gösterilmiştir. Dairelerin kesişim alanları dışındaki kısımlar bireylerin, görevlerin ve puanlayıcıların değişkenlik kaynaklarını, kesişim alanları ise birey-görev, birey-puanlayıcı ve görev-puanlayıcı ortak değişkenlik kaynaklarını ifade etmektedir. Bununla birlikte, dairelerin ortak kesişim alanı birey-görev-puanlayıcı ortak değişkenlik kaynağı ve hata değişkenlik kaynağını göstermektedir. Bu desen ayrıca bu araştırmada kullanılan desendir. Tümüyle çaprazlanmış desenler, veri toplamanın en bilinen yoludur, ancak pratik nedenlerden dolayı gerekli olabilecek başka seçenekler de mevcuttur (Strube, 2000).

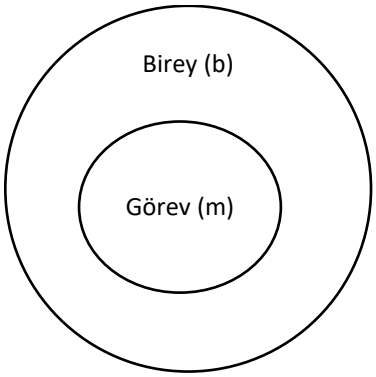
Kabul edilebilir gözlemler evreninde her zaman tüm yüzeylerin tamamen çaprazlanmış desen olarak bulunmadığı durumlar da mevcuttur. Örneğin, birçok testin içerik kategorileri *(h)* içine yuvalanmış maddelerden *(i)* oluştuğu görülebilir. Bu gibi durumlarda, kabul edilebilir gözlemler evrenini, evrendeki her maddenin *(i)* yalnızca bir içerik kategorisi *(h)* ile ilişkilendirildiği *i:h* şeklinde ifade etmek doğru olacaktır (Brennan 2001). Çaprazlanmış desenin yanında, farklı bireylere *(b)* farklı maddelerin *(m)* sunulduğu ve farklı puanlayıcıların *(p)* farklı öğrencilerin cevaplarını puanladığı bir desen oluşturulduğunda buna *tümüyle yuvalanmış (fully-nested) desen* denilmektedir (Shavelson ve Webb, 1991). Bu desen, *:* işaretinin yüzeylerin arasına konması şeklinde gösterilir. Yukarıda bulunan yazma testi hatırlanacak olursa, testi alan bireylerin yanıtlarını aynı değil de farklı puanlayıcıların puanladığı bir senaryo *px(m:p)* şeklinde olacaktır (Brennan 2001).

Tablo 5

Yuvalanmış Desene İlişkin Örnek (px(m:p))

	Madde1		Madde2		Madde3		Madde4	
Birey	Puanlayıcı 1	Puanlayıcı 2	Puanlayıcı 3	Puanlayıcı 4	Puanlayıcı 5	Puanlayıcı 6	Puanlayıcı 7	Puanlayıcı 8
1	2	3	4	4	2	2	3	3
2	2	3	1	4	3	3	1	3
3	1	2	3	1	2	3	3	2
4	1	1	4	3	2	2	3	1

Tek yüzeyli bir evrende m:b yuvalanmış desen için venn şeması Şekil 3'te gösterilmiştir.



Şekil 3. Tek Yüzeyli m:b yuvalanmış desen modeli

Dengelenmiş ve dengelenmemiş desenler. G ve D çalışmaları yapılırken ayrıca veri niteliklerine göre dengelenmiş ve dengelenmemiş (balanced ve unbalanced) olmak üzere farklı desenler de bulunmaktadır. Dengelenmiş desende eksik veri yoktur ve yuvalanmış herhangi bir yüzey için gözlem sayısı o yüzeyin her seviyesi için sabittir. Yüzeye ait koşullardaki gözlem sayısının farklılaşması durumunda ise verinin dengelenmemiş olduğu çıkarımı yapılır (Brennan, 2001). Bu duruma bir örnek; birbirinden farklı okulların öğrencilerinin bulunduğu bir çalışmada bireyler (b) ve okul (o) yüzeylerinin yuvalanmış b:o deseninde, çalışmada yer alan her okula ait öğrencilerin sayısı eşit olduğunda bu verinin dengelenmiş (balanced), öğrencilerin sayılarının eşit olmadığı durumda ise verinin dengelenmemiş (unbalanced) olduğu söylenebilir.

Varyans bileşenleri. Varyans bileşenleri, bir ölçme işlemine verilen yanıt arasındaki farklı değişkenlik kaynaklarını ifade eder. Başka bir deyişle, sınava giren kişiler için gözlenen puanlar arasındaki farklar çeşitli nedenlerden kaynaklanabilir. Örneğin, bireyin bir görevden aldığı puan birey etkisi (b) (ki bu aynı zamanda ölçmenin objesidir), bireylerin tutumuna veya yeteneklerine bağlı sistematik değişkenler, iyi veya kötü yazılmış maddelerden kaynaklanan madde etkisi (m) ve birey-madde etkileşimi (bm) gibi değişkenlik kaynaklarından söz edilebilir. Bu nedenle, bir bireyin bir görevden aldığı gözlenen puan şu şekilde ifade edilebilir:

$$X_{bm} = \mu + (\mu_b - \mu) + (\mu_m - \mu) + (X_{bm} - \mu_b - \mu_m + \mu)$$

X_{bm} Evrendeki herhangi bir maddede herhangi bir birey için gözlenen puanı ifade ettiğinde, μ tüm bireyler ve maddeler üzerindeki genel ortalama puan, μ_b birey (b) için tüm maddeler üzerindeki ortalama puan, μ_m (m) maddesi için tüm bireyler üzerindeki ortalama puandır. $(\mu_b - \mu)$ Birey etkisini temsil ederken, $\mu_m - \mu$ madde etkisidir ve $(X_{bm} - \mu_b - \mu_m + \mu)$ desende belirtilmeyen birey-madde etkileşimini ve diğer tüm hata kaynaklarını içeren artık *(residual)* etkidir.

Bağıl ve mutlak hata varyansı. G kuramında, bağıl ve mutlak değerlendirmelere dayalı kararlara karşılık gelen iki tür hata varyansı ayrı olarak tanımlanmaktadır. Bağıl kararlar, öğrenciler arasındaki bireysel farklılıklar hakkındaki kararlardır. Mutlak kararlar ise mutlak performans düzeyi ile ilgili kararlardır (Shavelson ve Webb, 1991).

Bazı durumlarda, özellikle ölçüt referanslı değerlendirme alanlarında, bir öğrencinin önceden belirlenmiş bir seviyede performans gösterip gösteremeyeceğine dair kararlar alınabilir. Bu durumlarda, *mutlak hata varyansı* ile ilgilenilir (Brennan, 2001). Bu hem öğrencilerin sıralamasıyla ilgili bilgileri hem de ortalama puanlardaki farklılıkları yansıtır (Shavelson ve Webb, 1991). Bireylere ait gözlenen ve evren puanlarının arasındaki farka ilişkin oluşan varyansa *mutlak hata varyansı* adı verilir ve $\sigma^2(\Delta)$ ifadesi ile gösterilir. Bu varyans ölçme objesine yönelik olan evren puan varyansı hariç diğer bütün varyansların toplamı şeklindedir. Mutlak değerlendirmelere dayalı kararlarda kullanıldığı için bireylerdeki değişimin gruba bağlı olmadan belirlenmesi amacıyla mutlak bir ölçüt kullanılarak elde edilir (Brennan, 2001). Ölçme objesi dışındaki tüm kaynaklar mutlak kararlar için bir hata kaynağıdır (Strube, 2000). Bu nedenle, tek yönlü (b × m) desen örneğinde, mutlak hata varyansı, hem madde etkisine hem de ölçmede kullanılan madde sayısı üzerinden ortalaması alınan artık etkiye *(resudial)* bağlı varyans bileşenlerini içerir. Bireyler (b) için mutlak hata varyansı:

$$\Delta_b \equiv \overline{X}_b - \mu_b$$

$$= \nu_m + \nu_{pm} \qquad (14)$$

Mutlak hata, genellikle ölçümlerin ölçüt referanslı yorumlarıyla ilişkilendirilir. Mutlak hata varyansı eşitlik X'ten yola çıkılarak şu şekilde gösterilir (Brennan, 2003):

$$\sigma^2(\Delta) = \sigma^2(m) + \sigma^2(bm) = \sigma^2(m)/n'_m + \sigma^2(bm)/n'_m \quad (15)$$

Bağıl hata varyansı ise, bireylere ait gözlenen ve evren puanlarının arasındaki farklılığın büyüklüğüne yönelik bulunan hata varyansıdır. Burada gözlenen ve evrene ait olan puanlar popülasyon ortalamaları dikkate alınarak hazırlandığı için bağıl olarak anılır (Shavelson ve Webb, 1991). Araştırmacılar bireylerin birbirleri arasındaki sıralamalarını içeren kararlarla ilgilendiğinde öncelikli olarak bağıl hata varyansı ile ilgilenirler. Böyle bir durumda hata kaynakları, bireylerin ölçme koşullarının rastgele örneklenmesi ile oluşturulan yüzeylerle etkileşimleri ile sınırlıdır. Bunun nedeni, ölçme objesini içeren etkileşimlerin, yüzey düzeylerindeki bağıl değişiklikleri yansıtmasıdır (Brennan, 2001; Shavelson ve Webb, 1991). Bağıl hata varyansının kestirimi, ölçme işleminde kullanılan madde sayısı üzerinden kalan varyansın ortalaması alınarak bulunabilir (Brennan, 2001).Bağıl hata varyansı $\sigma^2(\delta)$ şeklinde ifade edilmektedir.

Bağıl değerlendirmelere dayalı yorumlar yapabilmek için, bireyin ham puanı olan X_{bM} anlamlı bir bilgi vermez. Bu anlamlı bilgiyi verecek olan bireyin sapma puanı, yani $X_{bm} - \mu_m$'dir. Bu sapma puanı, bireyin evren sapma puanının bir tahmini olan $\mu_b - \mu$ şekline yorumlanır. Bu mantık doğrultusunda, *b* birey için bağıl hata eşitliği şu şekilde gösterilir (Brennan, 2003):

$$\delta_b \equiv \left(\overline{X} - E_b\,\overline{X}_b\right) - \left(\mu_b - E_b\,\mu_b\right) \quad (16)$$

$$= \left(\overline{X}_b - \mu_m\right) - \left(\mu_b - \mu\right) = \nu_{bm}. \quad (17)$$

Buradan hareketle, bağıl hata varyansı:

$$\sigma^2(\delta) = \sigma^2(bm) = \sigma^2(bm)/n'_i. \quad (18)$$

şeklinde gösterilebilir. Bağıl hata varyansı, klasik test teorisindeki *hata varyansına* karşılık gelirken, mutlak hata varyansı *"genel" hata varyansı (generic error variance)* ile ilgilidir.

Genellenebilirlik katsayısı. Cronbach ve arkadaşları (1972) KTK'daki güvenirlik katsayısına benzer şekilde $E\rho^2$ olarak gösterilen ve G katsayısı adı verilen bir güvenirlik katsayı tanımlamıştır. G katsayısı, evren puan varyansının beklenen gözlenen puan varyansına oranı olarak görülebilir. (Brennan, 2003). Bir yüzeyli bxp desenindeki bir G çalışmasında G katsayısı:

$$E\rho^2 = \frac{\sigma^2(b)}{\sigma^2(b) + \sigma^2(bp)} \tag{19}$$

Şeklinde ifade edilir. Burada $E\rho^2$ G katsayısını, $\sigma^2(b)$ gerçek puan varyansını ve $\sigma^2(bP)$ ise bireyin puanlayıcı etkileşimine göre varyansını ifade etmektedir. $E\rho^2$ Katsayısı, güvenirlik katsayısı olarak yorumlanabilir ve kullanılabilir. $n_i = n_i^{'}$ Olduğunda aslında bu katsayı Cronbach'ın Alpha güvenirliğine eşittir (Brennan, 2006). G katsayısı sıfır ile bir aralığında bir değer alır. Tek yüzeyli b * P rastgele deseninde, evren puan varyansı, ölçme objesi (σ_b^2) için varyans bileşenidir. Ölçme objesi hata varyansı yaratmadığı için bir yüzey olarak kabul edilmez. Sadece bunun dışında kalan varyans bileşenlerindeki σ^2_{bP}, ölçme objesi (bireyler) ile ilişkilidir ve bu nedenle, bağıl hata varyansı $\sigma^2(\delta)$ olarak sınıflandırılır. G katsayısı yüksek olduğunda, G çalışmalarındaki yüzeylerin diğer durumlara genellenebilmesi için güvenilir olduğu kabul edilir.

G kuramı mevcut tasarımda, KTK'da bulunmayan bir güvenirlik (dependability) indeksi de sunar. Mutlak güvenirlik katsayısı olan güvenirlik (dependability) indeksi, evren puan varyansını Gözlenen puan varyansının beklenen değerine oranlanmasıyla bulunur. Güvenirlik (dependability) Endeksi (phi katsayısı), aynı zamanda evren puan varyansının gözlenen puan varyansına oranı olduğu için G katsayısına benzer. Ancak buradaki gözlenen puan varyansı, evren

puan varyansı ve mutlak hata varyansının bir tür birleşimidir ve eşitlik 20'deki gibi gösterilir:

$$\varphi = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_\Delta^2} \tag{20}$$

Burada "φ" güvenirlik (dependability) indeksini, "σ_b^2" evren puan varyansını ve "σ_Δ^2" ise mutlak hata varyansını ifade etmektedir. Mutlak hata varyansı, tüm hata varyansı kaynaklarını içerir, ancak evren puan varyansını içermez. Bu katsayı bireylerin performansını belirlenmiş standartlara göre kestirir ve eşitlik 20'den yola çıkılarak şu şekilde gösterilir:

$$\varphi = \frac{\sigma^2(b)}{\sigma^2(b) + \sigma^2(p) + \sigma^2(bp)} \tag{21}$$

KTK altında güvenirlik katsayılarının teori ve uygulamaları ve G kuramı altındaki $E\rho^2$ ve φ katsayılarının tümü, rastgele örnekleme varsayımına bağlıdır. Ölçme nesneleri olan bireylerin, popülasyondan rastgele bir örneklem olduğu varsayılır. Varyans bileşenlerinin kestirimi aynı zamanda, her yüzeyin seviyelerinin tüm olası koşullar arasından rastgele örneklendiği varsayımına bağlıdır. Klasik test teorisi ve varyans analizinin sadece basit bir birleşiminin ötesinde, G kuramı, benzersiz kavramsal çerçevesi altında; farklı hataları ayırt etmede, varyans bileşenlerini detaylandırmada ve kabul edilebilir gözlemler evreni olan G çalışmalarının yanı sıra genelleme evreni olan D çalışmalarında da uygun tahminler yapmada nispeten avantajlıdır.

G katsayısı ($E\rho^2$), bireylerin bağıl sıralamasını ilgilendiren kararla ilişkilidir. Örneğin, ilk üç başvuru sahibinin alınacağı bir iş pozisyonu puanların bağıl yorumlanmasına dayanan bir karardır. Bu tür bir karar için genellikle norm referanslı puanlar kullanılır. Güvenirlik endeksi (φ) ise, diğer bireylerin performansından bağımsız olarak, bir bireyin performansının kesin düzeyine odaklanan mutlak bir kararla ilgilidir. Yüksek lisans eğitimi başvurusu için istenen 70 Yabancı Dil Puanı sınavı için belirlenen sabit kesme puanları buna örnek olarak gösterilebilir. Bu tür kararlar için kriter referanslı puanlar kullanılır. G katsayısının ve güvenirlik indeksinin doğru hesaplanması, ölçme işlemindeki bağıl ve mutlak hatanın doğru belirlenmesi ve sınıflandırılmasına bağlıdır. Genel olarak, mutlak hata varyansı, bağıl hata

varyansından daha büyük veya eşittir ve bu nedenle karşılık gelen güvenirlik indeksi, G katsayısından daha az veya eşit olmaktadır (Brennan, 2006).

İki yüzeyli çaprazlanmış b x m x p rastgele desen ile G modeli. Bu araştırmada kullanılan desen olan birey, madde ve puanlayıcının üç yüzey olduğu bir ölçme deseninde, *"b x m x p"* gibi iki yüzeyli dengeli bir desen oluşturulur. Bireyler, ölçme objesi olduğu için bir boyut gibi görülmez. Bu desene ilişkin ölçümlerin (X) varyansı şu şekilde gösterilir:

$$\sigma^2(X_{bmp}) = \sigma^2{}_b + \sigma^2{}_m + \sigma^2{}_p + \sigma^2{}_{bm} + \sigma^2{}_{bp} + \sigma^2{}_{mp} + \sigma^2{}_{bmp,e} \quad (22)$$

Eşitlik 22'de "$\sigma^2{}_b$" evren puanını temsil etmek üzere;

Bağıl hata varyansı, evren puan varyansı yanında sadece birey ile ilgili olan varyansların toplamıdır: $\sigma^2{}_{bm} + \sigma^2{}_{bp} + \sigma^2{}_{mp} + \sigma^2{}_{bmp}$.

Mutlak hata varyansı ise, evren puan varyansı dışındaki tüm varyans bileşenlerinin toplamıdır: $\sigma^2{}_m + \sigma^2{}_p + \sigma^2{}_{bm} + \sigma^2{}_{bp} + \sigma^2{}_{mp} + \sigma^2{}_{bmp}$.

Buradan hareketle *b x M x P* iki-yüzeyli dengelenmiş desende G katsayısı ($E\rho^2$):

$$E\rho^2 = \frac{\sigma_b^2}{\sigma_b^2 + \sigma^2{}_{bm} + \sigma^2{}_{bp} + \sigma^2{}_{bmp}} \quad (23)$$

Yine *b x M x P* iki yüzeyli dengelenmiş desenli güvenirlik indeksi phi (φ):

$$\varphi = \frac{\sigma_b^2}{\sigma^2{}_b + \sigma^2{}_m + \sigma^2{}_p + \sigma^2{}_{bm} + \sigma^2{}_{bp} + \sigma^2{}_{mp} + \sigma^2{}_{bmp}} \quad (24)$$

Yazılı anlatım becerilerinin ölçülmesi. Eğitim sistemleri pek çok farklı değişkenden etkilenen bir süreç sonunda temel girdi olan bireyleri en iyi şekilde işleme, hayata hazırlamayı, gerçek yaşamda hem kendileri hem de ülkeleri için faydalı bireyler hâline getirme amacını gütmektedir. Süreç içerisinde yapılan izleme ve değerlendirme çalışmaları bu hedeflere ne ölçüde yaklaşıldığını belirlemeye yardımcı olmaktadır. Bu bağlamda, örneğin Türkiye'de, farklı eğitim kademelerinde yapılan izleme amaçlı ulusal ve uluslararası sınavlardan elde edilen veriler öğrencilerin farklı alanlardaki bilgi ve beceri düzeylerinin anılan hedeflere yaklaşıp

yaklaşmadığını göstermesi bakımından değerli görülmektedir. Örneğin Millî Eğitim Bakanlığı (MEB) Ölçme Değerlendirme ve Sınav Hizmetleri Genel Müdürlüğünce (ÖDSGM) hayata geçirilmiş olan Akademik Becerilerin İzlenmesi ve Değerlendirilmesi (ABİDE) sınavıyla, ulusal düzeyde bireylerin üst düzey düşünme becerilerine sahip olma durumlarının; okuma, matematik ve fen alanlarındaki bilgi ve beceri düzeylerinin ve bunları etkileyen faktörlerin uzun vadede izlenmesi hedeflenmektedir (MEB ÖDSGM, 2016). Yine uluslararası düzeyde yapılan PISA ve TIMSS sınavlarından elde edilen veriler de uluslararası bir karşılaştırma yapmanın yanında bireylerin bulundukları sistem içinde ne kadar ilerleme kaydettiğine dair ipuçları sağlamaktadır. Türkiye'de uygulanan ulusal ve uluslararası izleme sınavları kapsam bakımından ele alındığında matematik ve fen alanlarının yanında okuma becerisine de yer verildiği görülmektedir. Ancak sınavlarda sadece okumaya odaklanılması gerek akademik gerekse kişisel ve sosyal gelişimin önemli bir göstergesi olan dili kullanma becerisinin çok sınırlı düzeyde değerlendirilmesiyle sonuçlanmaktadır. Her ne kadar okuduğunu anlamaya ilişkin düzey diğer beceriler için sınırlı bir kestirim yapma şansı sağlasa da gerçek seviyeyi belirleme konusunda sınırlı kalmaktadır.

4 Temel dil becerisinden biri olan yazma becerisi Özdemir(1987) tarafından bireylerin düşüncelerini ve konuşmalarını yazılı olarak anlatmaları olarak tanımlanmıştır. Yazma sembollerin öğrenilmesiyle başlayıp bu sembollerle içerik oluşturma ve bu içeriğin kelime, cümle, paragraf ve metin gibi uzayan bir örgüde devam etmesi süreci olarak da görülebilir. Yazma becerisinde doğası gereği içerik önemli bir yer tutar. Bireyin tüm yazılı ortamlarda kendini istenilen şekilde ifade edebilmesi yazılı anlatım becerisinin düzeyiyle doğrudan ilişkilidir. Dolayısıyla söz konusu becerinin geliştirilmesi büyük önem arz etmektedir. Yazılı anlatım ürünlerinin niteliği pek çok unsurla doğrudan bir ilişki içerisindedir. Öğretmenlerin öğretim becerileri, mevcut öğretim programı, eğitim materyalleri ve eğitim ortamı bu unsurlardan bazılarıdır. Bununla birlikte yazma becerisi bilişsel gelişim ile de ilişkilendirilmektedir.

İlkokul birinci sınıf itibariyle, bireyin yazılı dil işaretlerini sesli şekilde okuduğu ve söyleneni veya kendi düşüncesini yazabildiği evreyi belirtmek için "ilk okuma-yazma" terimi kullanılmaktadır. Sözlü anlatımdan yazılı anlatıma geçiş döneminden sessiz okumaya kadar olan evreyi kapsayan bu dönem, formal ana-dili öğretiminin

ilk aşaması olarak ifade edilmiştir (Ferah Özcan, 2007). Yazılı anlatım becerisini kazanan bireyin eğitim süreci boyunca programlar doğrultusunda bu becerisini hem içerik olarak hem de anlam bakımından sürekli geliştirmesi hedeflenmektedir. Sadece okul yaşantısı ile sınırlı olmayan bu beceri iş yaşamında iyi yerlere gelmek ve günlük hayata adapte olmak amacıyla da birey tarafından sıklıkla kullanılır. Günümüzde bireyin hayatında bu kadar önemli bir yer tutmasından dolayı yazma becerisi okuma ve hesaplama becerisiyle birlikte hayat boyu öğrenme kapsamında anahtar beceri olarak yer almıştır (Akbaş ve Özdemir, 2002).

Okul çağı ile birlikte öğrenilen yazma bireylerin düşüncelerini ifade edecekleri, kaydedecekleri ve dönüştürebilecekleri en önemli iletişim araçlarından biri olup ilerleyen yıllarda yaşamlarını farklı yönlerden etkilemektedir (Dennis ve Swinth, 2001; Hamstra-Bletz ve Blote, 1993; Tseng ve Cermak, 1993). Yazma becerisindeki gelişimin en belirgin etkisi dilin diğer beceri alanlarındaki gelişimde kendini göstermektedir. İlgili literatürde yazma ve okuma becerilerinin hemen hemen aynı zihinsel süreçleri ve aynı dilsel yapıları ortak kullandıkları, aralarında birbirini şekillendirecek kadar yüksek düzeyli bir ilişki olduğu dile getirilmektedir (Brown, Ignham ve Roen, 1986; Carillo, 2016; Knoeller, 2003; Ediger, 2002; Tierney ve Shanahan, 1991; Parodi, 2013). Bireylerin yazarken düşünmeye ve dil ile deneyim kazanmaya yönelmesi, aynı zihinsel süreçleri üretim veya çözümleme için kullanması gibi sebeplerle yazmada başarılı olan bireylerin diğer dil becerilerinde de daha başarılı ve tecrübeli hâle geldikleri belirtilmektedir (Daniels, Hyde ve Zemelman, 1998; Waring, 2007). Alanda yapılan uygulamalı çalışmalarda elde edilen sonuçların da bu doğrultuda olduğu görülmektedir (Aihara vd., 1999; Brown, Ignham ve Roen, 1986; Carillo, 2016; Collins ve Lee, 2005; Farahzad ve Emam, 2010; Paesani, 2016; Tunks, 2011). Yazma becerisindeki gelişim ve başarının etkili olduğu alanlar sadece dil ile doğrudan ilişkili dersler değildir. Yapılan çalışmalarda yazmanın farklı alanlardaki öğrenmeler açısından hem etkili bir araç hem de geleceğe yönelik güçlü bir yordayıcı olduğu tespit edilmiştir. Araştırmalarda yazmanın ilkokulda üniversiteye kadar matematik, fen bilgisi, kimya, biyoloji, coğrafya, tarih ve hatta çeşitli meslek derslerinde dahi bir öğrenme aracı olarak kullanıldığı (Bangert-Drowns, Hurley ve Wilkinson, 2004); bu derslerdeki başarının yordanmasında en güçlü yordayıcı olduğu ifade edilmektedir (Manfra vd., 2016; Preiss, Castillo, Grigorenko ve Manzi, 2013). Yazma becerisini akademik yaşam

açısından önemli kılan bir diğer husus ölçme değerlendirme süreçlerinde kullanılan yaklaşımlardır. Açık uçlu sorularla yapılan ulusal ve uluslararası sınavlar (PISA, TIMMS, ABİDE vb.) ile okullardaki yerel sınavlarda yazma farklı alanlardaki bilgi düzeyini göstermede kullanılan temel araç olarak karşımıza çıkmaktadır. Özellikle okullarda yapılan sınavların önemli bir kısmı, öğrencilerin, açık uçlu sorulara farklı düzeylerde (cümle, paragraf ya da metin) yazılı cevaplar vermesini gerektirmektedir. Öğrencilerin bu sınavlarda başarı gösterebilmesi, doğru cevabı bilmesi kadar bu cevabı en uygun şekilde yazıya aktarabilmesine de bağlıdır (Broomley, 1999; 2007). Ayrıca, özellikle son dönemlerde öğrencilerin gelişim sürecini ve gerçek performansını ortaya çıkarmada etkili sonuçlar ortaya koyan portfolyoların büyük ölçüde yazılı anlatım becerisini gerektirmesi de akademik başarı düzeyinin belirlenmesi açısından yazmayı ön plana çıkaran bir diğer gelişmedir (Al Khoudary, 2015; Waring, 2007). Yazmanın insan yaşamında etkili olduğu bir diğer önemli alan üst düzey düşünme becerilerinin gelişimidir. Yazma sürecinde bireylerin başarıları bilginin sınıflanması, sıralanması, mantıklı bir şekilde ilişkilendirilmesi, organize edilmesi, detaylandırılması veya özetlenmesi, eleştirel bakış açısıyla değerlendirilmesi, problem çözme ve yaratıcı düşünme yollarının kullanılarak işlenmesi gibi farklı düşünme becerilerinin kullanımına bağlı olarak artış göstermektedir (Bangert-Drowns vd., 2004). Dolayısıyla yazma, çok basit düzeyli bilişsel işlemlerden üst bilişsel işlemlere ve hatta üst düzey düşünme becerilerine kadar uzanan bütün süreçlerde düşünme yollarını somutlaştırması ve öğrencileri farklı şekillerde düşünmeye sevk etmesiyle önemli bir görevi yerine getirmektedir (Broomley, 1999). Son dönemde yapılan araştırmalar; basit düzeyli bilişsel işlemlerin ötesinde yaratıcı düşünme, eleştirel düşünme, mantıksal düşünme ve yansıtıcı düşünme (King, Goodson ve Rohani, 2009) gibi öğrenci başarısında oldukça etkili olan ve 21. yüzyıl becerileri olarak adlandırılan üst düzey düşünme becerilerinin hem öğretilmesinde hem de geliştirilmesinde yazmanın anahtar rolü üstlendiğini, bu beceriler ile yazma arasında karşılıklı ve çok güçlü bir ilişki olduğunu göstermektedir (AlKhoudary, 2015; Bangert-Drowns, vd. 2004; Etemadzadeha, Seifi ve Far, 2013; Faragher ve Huijser, 2014; Ganapathy ve Kaur, 2014; Grupta, Burke, Mehta, Greenbowe, 2015; Klimova, 2013; Leggette, McKim, Homeyer ve Rutherford, 2015; Manalo ve Sheppard, 2016; Pantaleo, 2017; Preiss, Castillo, Flotts ve San Martin, 2013).

Sonuç olarak yazmanın akademik gelişimde ve üst düzey düşünme becerilerinin gelişiminde önemli bir rolü olduğu ortadadır. Günümüzde yazma deneyiminin sınırları teknolojideki gelişmeler sayesinde geçmişte hiç olmadığı kadar genişlemiştir. Eskinden sadece okul hayatının bir parçası olarak yaşamımızın sınırlı bir bölümünde varlık gösteren yazma, bugün pek çok bireyin gündelik yaşamında kendine yer bulmuştur. İnternet yayıncılığı ve sosyal medya vasıtasıyla bundan on yıl önce yılda bir dilekçe yazmak veya bir alışveriş listesi hazırlamak dışında yazıya başvurmayan insanlar, artık farklı platformlardaki paylaşımlara veya haberlere yorumlar yazmakta, kendi durumlarını, duygu ve düşüncelerini yazılı olarak bildirmektedir. Bir başka ifade ile yazı geçmişte hiç olmadığı kadar gündelik yaşamımıza girmiş durumdadır (Johnson, 2017).

Öğrencilerin yazma becerilerinin ölçülmesi genellikle ulusal düzeyde yapılan ölçmelerin önemli bir parçasıdır. Yazma becerisinin ölçülmesi, öğrencilerin yazma yeterliğine ilişkin genel seviyesi hakkında bilgi edinmenin en geçerli yolu olarak kabul edilen bu değerlendirmelerdir. Bu, öğrencilerin "gerçek" yazma görevlerini yerine getirmeleri gerektiği anlamına gelir. Daha sonra öğrencilerin yazma performansları puanlayıcılar tarafından puanlanır. Bu nedenle yazma becerilerini ölçmek genellikle pek ekonomik değildir ve zaman alıcıdır. Ancak, çoktan seçmeli testlerle yazma becerilerinin değerlendirilmesi yerine, yazma görevlerinin değerlendirildiği bir süreci tercih etmek geçerli bir yöntem olarak kabul edilir (Schoonen, Vergeer ve Eiting, 1997).

Dereceli puanlama anahtarı. Yanıtı öğrenci tarafından oluşturulan sınavların puanlanmasındaki en temel sıkıntı puanlama sürecinde yaşanabilecek objektiflik sorunudur. Açık uçlu bir sınavda öğrenci yanıtlarının puanlayıcılar tarafından farklı şekilde puanlanması sınavın güvenirliğini olumsuz yönde etkileyebilecek bir durumdur. Bu durum öğrencinin sahip olduğu bilgi ve becerilerini bir arada kullanarak bir ürün oluşturduğu yazılı anlatım becerilerinin puanlamasında da söz konusudur.

Puanlama yöntemleri düşünüldüğünde, öğrencilerin farklı bilgi ve becerilerinin bir arada kullanılarak bir yanıt oluşturması istenen sorulardan oluşan testlerin (örneğin; yazılı anlatım becerilerinin ölçülmesi) puanlama süreci çoktan seçmeli sorular ile hazırlanmış olan sınavlara göre nispeten objektif olmama sorunuyla karşı karşıya kalabilir. Bu tip sınavlarda objektif puanlama yapılmasını

sağlayacak yöntemler bulunmaktadır. Dereceli puanlama anahtarları (DPA) bu amaç için kullanılabilecek araçlardan birisidir. Yazma becerileri literatürü incelendiğinde bütünsel DPA'lar ve analitik DPA'lar'ın daha çok kullanıldığı görülebilir (Weigle, 2002). Bu çalışmada bütünsel ve analtik DPA'lar kullanıldığı için özellikle bu iki araç üzerinde durulacaktır.

Yazılı Anlatım Becerilerinin Ölçülmesi ve Değerlendirilmesi Çalışması kapsamında görevlere göre uygunluğu da dikkate alınarak holistik (bütünsel) ve analitik dereceli puanlama anahtarları kullanılmıştır. Cümle görevlerinin puanlanmasında bütünsel, paragraf ve metin görevlerinin puanlanmasında ise analitik DPA kullanılmıştır.

Birçok değerlendirme sürecinde, bütünsel puanlama veya metnin genel izlenimine göre bir metne tek bir puanın verilmesi yöntemi kullanılır. Bütünsel puanlama son yıllarda sıklıkla kullanılmaktadır ve birçok olumlu özelliğe sahiptir. Bütünsel puanlamanın önde gelen savunucularından White'a (1984) göre, bütünsel puanlama, puanlayıcının dikkatini eksikliklere değil, yazılı ürünün güçlü yönlerine odaklamayı, böylece bireylerin yapamadıklarından ziyade iyi yaptıkları yönlerin ödüllendirilmesini amaçlamaktadır. Bütünsel DPA'lar, içerikte neyin en önemli olduğuna bağlı olarak puanlayıcıların dikkatini yazma görevinin belirli yönlerine odaklamak için tasarlanabilir ve böylece bu yönler hakkında etkili bir şekilde önemli bilgiler sağlayabilir (Weigle, 2002). Bütünsel DPA'larda öğrencinin performansı bir bütüne tek puan verme şeklindedir.

Analitik DPA'lar ile ölçülmek istenen özellikte bulunması beklenen yönleri içeren birden fazla ölçüt bulunur. Ölçmenin amacına bağlı olarak, örneğin bir yazma görevindeki ürün; içerik, organizasyon, anlam bütünlüğü, söz dağarcığı, dil bilgisi veya noktalama işaretleri gibi ölçütler bu tür DPA'larda bulunabilir. Her bir ölçüt kendi içinde bulunan ağırlıklandırmaya göre değerlendirilir ve bir toplam puan ortaya çıkar. Bu tür DPA'lar öğrenci performansının çeşitli boyutlarında bulunan başarı seviyeleri hakkında bilgi veren türden bir puanlama aracıdır. Bu sebeple öğrencilerin güçlü ve geliştirilmesi gereken yönlerine ilişkin bir çerçeve ortaya konulabilmektedir. Ölçülmek istenen özelliğe yönelik geri bildirim verilmek istendiğinde analitik DPA'ların kullanılmasının daha uygun olduğu değerlendirilmektedir. Bununla birlikte nispeten fazla sayıda ürünün değerlendirilmesi gerektiğinde analitik DPA'lar kullanışlı olmamakta ve puanlayıcıdan kaynaklanan hataların sürece etki etmesi

durumu sebebiyle güvenirliği olumsuz anlamda etkileyen etkenler söz konusu olabilmektedir. Öğrencilerin yazılı anlatım becerileri nitelikli bir şekilde ölçülmek istendiğinde iki bileşenden söz edilebilir. Bunlardan birincisi nitelikli hazırlanmış bir DPA ile puanlama yapılmasıdır. Diğeri ise DPA'nın kullanılmasında donanım ve deneyim sahibi puanlayıcıların bulunmasıdır. Weigle'ye (2002) göre bir DPA hazırlanırken dikkate alınması gereken hususlar şu şekildedir:

1. DPA'yı kim kullanacak?
2. Yazma görevinin hangi yönleri en önemli ve bunlar nasıl birbirinden ayrılacak?
3. Kaç tane puan veya puanlama düzeyi kullanılacak?
4. Puanlar nasıl raporlanacak?

İlgili Araştırmalar

Bu bölümde, puanlayıcı deneyimleri temelinde güvenirlik ile ilgili Türkiye'de ve dünyada yürütülen puanlayıcı güvenirliği çalışmalarına yer verilmiştir.

Shohamy, Gordon ve Kraemer (1992), yazma yeterliğini değerlendirmede profesyonel geçmişleri ve puanlama eğitimleri bakımından birbirinden farklı olan puanlayıcılar arasındaki puanlayıcı güvenirliğinin incelendiği araştırmalarında 12. Sınıf seviyesinde 250 EFL (Yabancı Dil Olarak İngilizce) öğrencisine altı görevden oluşan bir yazma görevi vermiştir. Çalışma için elli öğrencinin yanıtlarının bulunduğu bir örneklem belirlenmiştir. Örneklemin farklı seviyelerden yeterli yanıtlar içermesi için beş farklı yeterlik seviyesinden yanıtların bulunmasına özen gösterilmiştir. Puanlayıcılar ise profesyonel geçmişlerine (İngilizce öğretmeni olup olmama) ve puanlama eğitimlerine (bir grup yazılı örneklerin değerlendirildiği eğitim almışken diğer grubun almaması) göre seçilmiştir. Çalışma sonucunda, eğitimli puanlayıcılar için genel güvenirlik katsayıları, eğitimsiz puanlayıcılara göre daha yüksek bulunmuştur. Ayrıca, puanlama eğitimine sahip ve alan uzmanı olanlar için elde edilen yüksek puanlayıcı katsayıları, bu grup için puanlamanın zaman içinde sabit olduğunu göstermektedir. ANOVA sonuçları, eğitimin puanlamalar üzerinde önemli bir etkisi olduğunu göstermiştir.

Kan (2001), yazılı yoklama sınavlarına ilişkin puanlama durumu; puanlama cetveli ile yanıt anahtarı kullanımı, puanlamalar arasındaki süre, aynı ve farklı

puanlayıcılar tarafından gerçekleştirilen puanlama sürecinin güvenirliğine etkisini incelemiştir. Bu kapsamda 14 puanlayıcı ile görüşmeler yapılmış ve 9 yazma görevinden oluşan bir testin puanlanması istenmiştir. Çalışma sonucunda bazı öğretmenlerin kendi içlerinde, puanlama cetveli ve yanıt anahtarı kullanarak ve kullanmadan tutarlı puanlar vermedikleri gözlenmiştir. Bununlar birlikte, puanlamaların arasındaki farklılaşmanın ve puanlamalar arasındaki sürenin, aynı puanlayıcının kendi içinde yaptığı puanlamanın tutarlılığını olumsuz anlamda etkileyecek derecede bir etkiye sahip olduğu sonucuna ulaşılmıştır. Ulaşılan bir diğer sonuç ise, puanlama cetvelinin kullanılıp kullanılmama durumuna göre yapılan puanlamaların ortalamalar arasındaki tutatırlılığı bozacak derecede farklılığın olduğudur.

Barkaoui (2008), puanlama yönteminin ve puanlayıcı deneyiminin ESL (İkinci yabancı dil olarak İngilizce) yazma sınavı üzerindeki etkisini inceleyen araştırmasında deneyimsiz puanlayıcıların puanlayıcılar arasında ve kendi içlerinde daha fazla değişkenlik sergiledikleri, puanlama anahtarına daha sık başvurma, yazmanın kısmi özelliklerine daha sık odaklanma ve deneyimli puanlayıcılara göre metni yorumlamaya veya düzenlemeye daha fazla zaman ayırma eğiliminde olduğu sonucuna ulaşmıştır. Deneyimli puanlayıcılar ise puanlama anahtarındaki kriterlerin dışındaki başka kriterlere atıfta bulunma, daha fazla değerlendirme stratejisi kullanma, yazma görevlerini bir bütün halinde okumak ve değerlendirmek için daha fazla zaman harcama, daha verimli, kendinden emin, kendi içinde tutarlı olma eğilimi göstermiş ve deneyimsiz puanlayıcılara göre daha homojen puanlama gerçekleştirmişlerdir.

Derkuş (2009), çalışmasında problem çözme becerisine yönelik bir ölçek kullanarak puanlayıcılar-arası uzlaşmayı, sınıf içi korelasyon katsayısı ile lojistik regresyon tekniklerini kullanarak incelemiştir. Sekiz puanlayıcı tarafından puanlanan matematiksel problem çözme testinden sınıf içi korelasyon katsayısı ve lojistik regresyon sonucu elde edilen sıralamalar sonucuna göre iki teknik arasındaki sıra farkları korelasyon katsayısı -0,23 bulunmuş ve iki ayrı teknik sonucu elde edilen sıralamalar birbirleriyle düşük ve ters yönlü uyum göstermişlerdir.

Lim (2011), çalışmasında uluslararası bir İngilizce yeterlilik sınavı olan Michigan English Language Assessment Battery-MELAB'ın yazma bölümünden alınan hazır verileri kullanarak çok yönlü Rasch modeline dayalı ve 12 ila 21 aylık

boylamsal bir sürede yeni ve deneyimli puanlayıcıların puanlama kalitesini araştırmıştır. Çalışma sonuçlarına göre, deneyimsiz puanlayıcıların katılık ve tutarlılığının bazı durumlarda deneyimli meslektaşlarına göre ayırt edilemediği gözlenmiş ve ayırt edilebilir oldukları durumlarda, puanlama kalitesi daima daha kötü bulunmuştur. Bununla birlikte deneyimsiz puanlayıcıların puanlama kalitesinin süreç boyunca nispeten hızlı bir şekilde ilerlemesi de söz konusu olmuştur. Puanlayıcılar süreç boyunca kabul edilebilir puanlama kalite sınırları içinde kalmıştır.

Bıkmaz (2011), çalışmasında DPA türü ve puanlayıcı sayısının puanlayıcı güvenirliğine etkisini Kappa istatistiği, Loglinear analizi ve Krippendorff'un Alpha tekniğine göre incelemiştir. Çalışmada bütünsel ve analitik DPA'lar ile 2, 5 ve 10 puanlayıcıdan oluşan farklı puanlama durumları denenmiştir. Bu kapsamda analitik DPA'ların daha güvenilir puanlama sonuçları sağladığı, puanlayıcı sayılarının artırılmasının her üç teknik sonucuna göre puanlayıcı güvenirliğini düşürdüğü ve en yüksek puanlayıcı güvenirliğinin 2 puanlayıcı ile yapılan puanlama durumu sonucu elde edildiği gözlenmiştir. Bunun yanında; Krippendorff'un Alpha ve Kappa tekniklerinde sonuçların benzer olduğu, Krippendorff'un Alpha tekniğinin puanlayıcı sayısı değişiminden Kappa tekniğine göre daha az etkilendiği ve Loglinear analizi tekniğinin değişkenler arasındaki etkileşim ve uyumsuzluk kaynaklarını göstermesi bakımından daha kapsamlı bilgi verdiğine de araştırma sonucunda ulaşılmıştır.

Büyükkıdık'ın (2012), problem çözme becerilerinin değerlendirildiği test durumuna yönelik yaptığı araştırmasında iki performans görevi dört puanlayıcı tarafından analitik ve bütünsel DPA'lar ile puanlanmıştır. Puanlama sonucunda elde edilen veriler KTK ve G kuramı ile puanlayıcı güvenirliği bağlamında incelenmiştir. Araştırma sonunda G kuramı ile elde edilen katsayıların KTK'ya dayalı yöntemlerle elde edilen katsayılara oranla nispeten daha yüksek olduğu ve G kuramı analizleri sonucu elde edilen sonuçların daha detaylı bilgi verdiği sonucuna ulaşılmıştır. Araştırmada ayrıca analitik DPA'lardan elde edilen puanların nispeten daha yüksek güvenirliğe sahip olduğu ifade edilmiştir.

Çakıcı Eser ve Gelbal (2013), araştırmalarında PISA sınav sorularından yararlanılarak oluşturulan bir test formuna dayalı yapılan ölçme işleminin sonucu 3 puanlayıcı tarafından puanlanmış ve sonuçları G kuramı ile lojistik regresyon analizine göre incelemiştir. Araştırma sonucunda her iki tekniğe göre benzer

sonuçlara ulaşıldığı ancak lojistik regresyon analizine dayalı çıktıların G kuramı kadar duyarlı sonuçlar vermediği gözlenmiştir. Bu bağlamda lojistik regresyon analizinin G kuramına göre daha yüzeysel sonuçlar sunduğu sonucuna ulaşılmıştır.

Han (2013), puanlama yönteminin ve puanlayıcı eğitiminin güvenirliğe etkisini inceleyen bir araştırma yapmıştır. İlgili araştırmada EFL kompozisyonları bütünsel ve analitik DPA'lar ile detaylı puanlama eğitimi almış ve bu alanda temel bir eğitim almış puanlayıcılar tarafından puanlanmıştır. Elde edilen sonuçlar analiz edildiğinde detaylı puanlama eğitim almış puanlayıcıların her iki DPA türünde de tutarlı puanlamalar yaptıkları, sadece temel bir eğitim almış olan puanlayıcıların ise düşük nitelikte puanlama yaptıkları ve her iki DPA türünde oldukça farklılaştıkları sonucuna ulaşılmıştır. Bu bağlamda puanlama eğitiminin puanlayıcı davranışlarına olumlu anlamda önemli katkıları olduğu da araştırmanın bir diğer çıktısıdır.

Yıldıztekin (2014), araştırmasında 7. sınıf öğrencilerine matematikte problem çözme becerilerini ölçen açık uçlu sorulardan oluşan bir test uygulamış ve sonuçlar analitik ve bütünsel DPA ile beş farklı matematik öğretmeni tarafından puanlanmıştır. Elde edilen puanların KTK ve G kuramına göre güvenirlik kestirimleri yapılmış ve puanlayıcılar arasındaki tutarlılık derecesi belirlenmeye çalışılmıştır. KTK kapsamında Pearson ve Spearman korelasyonu, Cronbach'ın Alpha, Kappa ve Krippendorff'un Alpha katsayısı tekniklerinin kullanıldığı araştırma sonuçlarına göre hem KTK hem G kuramı kapsamında sonuçların Kappa istatistiği dışında birbirine paralel ve yüksek olduğu sonucuna ulaşılmıştır. Bunun yanında analitik DPA'lar ile gerçekleştirilen puanlamadan elde edilen uyum düzeyleri bütünsel DPA'lara göre nispeten daha yüksek bulunmuştur.

Attali (2015), kompozisyon yazma görevine verilen yanıtların değerlendirildiği araştırmasında yeni eğitim almış ve uzman puanlayıcılar arasında güvenirlik bağlamında manidar bir farklılık olup olmadığını belirlemeye çalışmıştır. İlgili araştırma kapsamında deneyimsiz puanlayıcılara doğru değerlendirmeler hakkında anında geri bildirim sağlanan kısa bir eğitim verilmiştir. Daha önce puanlama deneyimi çok az olan veya hiç olmayan 14 katılımcı bu eğitimi tamamlamış ve yeni eğitilen puanlayıcıların performansı, yazma görevine verilen yanıtları puanlama konusunda kapsamlı deneyime sahip 16 uzman puanlayıcının performansıyla karşılaştırılmıştır. Sonuçlar G kuramı kapsamında incelenmiş ve yeni eğitilmiş

puanlayıcı grubundan elde edilen puanların, deneyimli puanlayıcı grubundan alınan puanlarla benzer ölçüm özellikleri sergilediği gözlenmiştir.

Kara ve Kelecioğlu (2015), kesme puanlarının belirlenmesinde puanlayıcı etkisinin araştırıldığı çalışmalarında öğretmenler ve alan uzmanı olacak şekilde iki puanlayıcı grubunun puanlamalarını G kuramı kapsamında karşılaştırmış ve puanlayıcı yeterliklerinin kesme puanlarına etkisini değerlendirmeye çalışmışlardır. Bu kapsamda standart belirleme yöntemleri olan 1-0 ve Nedelsky yöntemlerini kullanarak verileri tek ve iki değişkenlik kaynaklı tümüyle çaprazlanmış desen şeklinde analiz etmişlerdir. Araştırma sonucuna göre, her iki yöntemde alan uzmanı grubun öğretmenlere kıyasla belirgin olmamakla birlikte daha tutarlı değerlendirmeler yaptığı gözlenmiş ve ilgili sonuçlar G kuramına dayalı G ve Phi katsayılarıyla desteklenmiştir.

Güler ve Taşdelen Teker (2015), araştırmasında puanlayıcılar arası güvenirliğin tahminine yönelik korelasyon, ortalamaların karşılaştırılması, uyum yüzdesi ve G kuramı olmak üzere dört tekniği karşılaştırmıştır. İki puanlayıcı tarafından 10 madde üzerinde gerçekleştirilen puanlama sonucunda kullanılan teknikler sonucunda en yüksek kestirim G kuramı ile elde edilmiştir (0,90). Ayrıca korelasyon katsayısı pozitif ve yüksek (0,74), uyum yüzdesi %58,9 olarak bulunmuş ve puan ortalamaları arasında istatistiksel olarak manidar bir fark bulunamamıştır. Araştırma sonucuna göre, hesaplanması nispeten daha karmaşık olmasına rağmen en detaylı bilginin G kuramı ile verilebileceği vurgulanmıştır.

Bıkmaz Bilgen ve Doğan (2017), puanlayıcı güvenirliğinin farklı tekniklerle karşılaştırıldığı bir araştırma yürütmüşlerdir. Bu kapsamda DPA türü ile puanlayıcı sayısındaki değişimin, puanlayıcı güvenirliği üzerindeki sonuçları farklı tekniklerle belirlenmeye çalışılmıştır. Betimsel niteliği olan ilgili araştırmada Kappa istatistiği, log linear analiz tekniği ve Krippendorff'un Alpha tekniği kullanılmıştır. Araştırma sonunda üç teknik ile elde edilen analiz sonuçlarına göre, analitik DPA kullanımı sonucu elde edilen puanlarda, puanlayıcı sayısının artmasıyla güvenirlik oranının düştüğü gözlenmiştir. Tüm teknik analizlerde en yüksek güvenirlik düzeyleri iki puanlayıcının kullanıldığı durumlarda gözlenmiştir. Kappa ile Krippendorff'un Alpha tekniği birbirine paralel sonuçlar vermiş ve Krippendorff'un Alpha tekniğinin puanlayıcı sayısındaki değişimden Kappa tekniğine oranla daha az etkilendiği gözlenmiştir. Bunun yanında Log-linear analiz tekniğinde değişkenler arası etkileşim

ve uyumsuzluk kaynaklarınının gösterildiği daha ayrıntılı ve kapsamlı bilgi sağlandığı sonucuna ulaşılmıştır.

Anadol (2017), Araştırmasında DPA'lar ile İngilizce yazılı anlatım becerilerinin üç grup tarafından puanlanmasıyla elde edilen verileri G kuramı kullanarak karşılaştırmıştır. Puanlayıcı grup, DPA kullanımına yönelik az deneyimli, çok deneyimli ve deneyimi az ve çok olanlar olmak üzere üç farklı grup şeklinde G çalışmaları yürütülmüştür. Araştırma sonuçlarına göre her üç grup için de varyans oranları, G ve Phi katsayıları ile mutlak ve bağıl hata varyanslarının paralellik gösterdiği sonucuna ulaşılmıştır. Buradan hareketle nitelikli bir DPA ile yapılan puanlamada deneyim yılının manidar bir etkisinin olmadığı sonucu elde edilmiştir.

Şahan (2018), araştırmasında puanlayıcı deneyiminin ve kompozisyon kalitesinin puanlayıcı davranışı ve puanlama üzerindeki etkisini incelemiştir. Puanlayıcıların düşük, orta ve yüksek deneyimli olmak üzere üç farklı deneyim grubundan seçildiği araştırmada analitik DPA'lar kullanılarak EFL kompozisyonları puanlanmış ve puanlayıcıların karar verme süreçleri kaydedilmiştir. G kuramı ile betimleyici ve yordayıcı istatistikler kullanılarak yapılan analizler sonucunda yüksek deneyimli puanlayıcıların öğrencilerin kompozisyonlarına karşı daha olumlu olduğunu ve daha az deneyimli eş puanlayıcılara kıyasla daha yüksek puanlar verdiği görülmüştür. Bunun yanında yüksek ve düşük deneyimli gruplar, toplam puan açısından da önemli ölçüde farklılık göstermiştir. Ek olarak, ölçme hatası kaynaklarını belirlemek için yapılan analiz sonuçlarında, yüksek ve düşük kaliteli kompozisyonlar bir bütün olarak ele alındığında nispeten daha küçük bir puanlayıcı etkisi sağlarken farklı kompozisyon kaliteleri için ayrı analizler yapıldığında puanlayıcıların farklılaşmaya daha fazla katkıda bulunduğu ifade edilmiştir. Ayrıca nitel bulgular, farklı deneyim gruplarındaki puanlayıcıların, farklı yeterlik seviyelerindeki kompozisyonları değerlendirirken farklı karar verme davranışları sergilediklerini göstermiştir.

Arslan Mancar'ın (2019), çalışmasında bağımsız puanlayıcılardan elde edilen puanları KTK ve G kuramı açısından uyum düzeylerini belirlemek amacıyla puanlayıcılar arasındaki uyum düzeyleri iki, üç ve beş puanlayıcı kullanılarak analitik DPA'lardan elde edilen puanların Kappa istatistiği, Krippendorff'un Alpha tekniği ve G kuramına yönelik analizleri ile incelenmiştir. Bu çalışmalara ek olarak, performansa dayalı değerlendirme süreci ve analitik DPA kullanımına ilişkin

puanlayıcı görüşlerine de yer verilmiştir. Araştırma bulgularına göre, performansa dayalı vaka çalışmalarının kapsamlı uygulamaları nedeniyle Kappa ve Krippendorff'un Alpha istatistiksel teknikleri yetersiz bulunmuştur. Bunun yanında çalışmada yapılan analizler sonucu G kuramına dayalı karar çalışmalarının kapsamlı bilgi ve ayrıntılı fikirler verdiğine ulaşılmıştır.

Gürlen, Boztunç Öztürk ve Eminoğlu (2019), ilkokul düzeyinde yapılan öğretmen, öz ve akran değerlendirmelerinin güvenirlik katsayılarını belirlemeyi amaçladıkları araştırmalarında Ankara ilinde bir devlet okulunda okuyan 30 üçüncü sınıf öğrencisi üzerinde incelemeler yapmışlardır. Çalışmanın amacı doğrultusunda verilerin analizinde G kuramı kullanılmıştır. Araştırma sonucunda öğrenci ana etkisi için tahmin edilen varyans bileşeninin her üç derste de toplam varyansın en büyük bileşeni olduğu tespit edilmiştir, G ve Φ katsayıları incelendiğinde müzik dersinde 0,80'in üzerinde, Türkçe ve sosyal bilgiler derslerinde ise 0,90'ın üzerinde güvenirlik katsayıları bulunmuştur. G yüzey analizi sonuçlarına göre, öğretmen ve akran değerlendirmeleri sırasıyla analizden çıkarıldığında G ve Φ katsayıları azalma eğilimi gösterirken, öz değerlendirme analizden çıkarıldığında bu katsayılar arttığı gözlenmiştir.

Carballo-Fazanes, Rey, Valentini, Rodriguez-Fernandez, Varela-Casal, Rico-Diaz, Barcala-Furelos ve Abelairaas-Gomez (2021), yaptıkları araştırmada Kaba Motor Gelişim Testi ile 3 ila 10 yaş arasındaki çocuklarda temel hareket becerilerini değerlendirmek amacıyla uzman ve acemi puanlayıcılar tarafından puanlama işlemi gerçekleştirilmiştir. Puanlayıcılar, her çocuğun performansını iki farklı şekilde (yavaş ve normal çekim video) izleyerek puanlamıştır. Puanlayıcılar arasındaki uyumu belirlemek için sınıf içi korelasyon katsayısı kullanılmıştır. Sonuçlara göre uzman puanlayıcılar ve beden eğitimi geçmişine sahip acemi puanlayıcılar bu alanda herhangi bir geçmiş deneyimi bulunmayan acemi puanlayıcılara göre daha yüksek puanlayıcı içi güvenirlik değerlerine ulaşmışlardır. Ancak, puanlayıcılar arası güvenirlik, deneyimlerine veya geçmişlerine bakılmaksızın tüm puanlayıcılarda nispeten değişken bulunmuştur.

Son yıllarda yürütülen araştırmalarda, puanlayıcı deneyimi ve geçmişinin puanlayıcı güvenirliğine etkisini belirleme amacıyla yapılmış değerlendirme çalışmaları yoğunlukla incelenmiştir. Puanlayıcı güvenirliği çalışmalarının genelde eğitim alıp almama, alan uzmanı olup olmama, puanlama deneyimi ve geçmişi ve

DPA kullanımına yönelik uzmanlık derecesi gibi değişkenlerle oluşturulan modellerin kullanıldığı gözlenmiştir. Güvenirlik çalışmalarında sıkça KTK'ya dayalı yöntemler ve değişkenlik kaynağına odaklanmayı sağlayan G kuramı analizleri kullanılmıştır. Araştırmaların sonuçlarında, puanlayıcı deneyimi ve uzmanlığı ile alana dair arka plana sahip olmanın uyum düzeyini ve dolayısıyla güvenirliği artırdığı görülmüştür. Ayrıca kullanılan DPA türü ile niteliğinin de puanlama güvenirliğinde önemli rol oynadığı sonucuna ulaşılmıştır. Bu sebeple, puanlama deneyiminin etkisine dair kanıt, karşılaştırma için dayanak ve daha sonra yapılacak araştırmalar için bilgi vermesi açısından yazılı anlatım becerilerinin puanlanmasında puanlayıcı deneyiminin güvenirliğe etkisinin farklı yöntemlerle incelenmesinin alana katkı sunması beklenmektedir.

Bölüm 3
Yöntem

Bu bölüm altında; araştırmanın türü, çalışma grubu, verilerin elde edilmesi, veri toplama araçlarının özellikleri ile verilerin analizine yer verilmiştir. Analiz sırasında kullanılan yazılımlar, puanlayıcı güvenirliklerine dair bazı varsayımlar ve bunlara uygun testler kullanılarak açıklanmıştır.

Araştırmanın Türü

Araştırmada, MEB ÖDSGM tarafından 2017-2018 yılları arasında uygulanan Yazılı Anlatım Becerilerinin Ölçülmesi ve Değerlendirilmesi Çalışması kapsamında yapılan Yazma Becerileri Testi'nin puanlama çalışmasından elde edilen sonuçlar kullanılarak puanlayıcı deneyimlerinin puanlamaya etkisini ortaya koymak için puanlayıcılar arası güvenirliğin Cohen'in ağırlıklandırılmış Kappa istatistiği, Krippendorff'un Alpha katsayısı ve genellenebilirlik kuramı bakımından ayrı ayrı hesaplanıp puanlayıcı güvenirliklerine ilişkin var olan durum incelendiği için ilişkisel ve betimsel bir araştırma olarak kabul edilebilir. Betimsel araştırma, belirli bir grup bireyin davranışlarını, düşüncelerini veya duygularını tanımlayan, olay ve durumların detaylı olarak açıklanması amacıyla yapılan araştırmalardır (Leary, 2014). Bununla birlikte, puanlayıcı deneyim etkisinin puanlama sonuçlarına bir etkisi olup olmadığı araştırılacağı için bu araştırma aynı zamanda yarı-deneysel araştırma olarak da kabul edilebilir. Çoğu durumda, araştırmacılar bağımsız değişkeni veya diğer faktörleri kontrol edemezler. Yarı-deneysel tasarımlar deneysel kontrol olanağı olmayan durumlar için kullanılmaktadır (Leary, 2014).

Araştırmanın Evreni ve Örneklemi

Araştırma kapsamında Adana, Ankara ve İstanbul illerinde MEB ÖDSGM tarafından 2017 yılında uygulanan "Yazma Becerileri Testi"ne katılan 9241 4, 7 ve 9. sınıf öğrencisi arasından seçkisiz olarak 240 öğrenci belirlenmiştir. Bu seçim yapılırken, analizlere uygun olması amacıyla, herhangi bir yanıtında kayıp verisi bulunan öğrencilerin verileri ayıklanmıştır. Araştırmanın örneklemi bu öğrencilerin cevaplarını puanlayacak olan Milli Eğitim Bakanlığı'na bağlı okullarda görev yapan ve daha önce yazma becerilerinin puanlanması çalışmasına katılan deneyimli

puanlayıcılar ile puanlama deneyimi olmayan puanlayıcılar arasından seçkisiz olarak belirlenmiş, her branşta (sınıf öğretmeni, Türkçe öğretmeni ve Türk dili ve edebiyatı öğretmeni) 4 olmak üzere toplam 12 öğretmenden oluşmaktadır. Tablo 6 puanlayıcıların branş ve deneyimlerine göre dağılımlarını göstermektedir.

Tablo 6.

Puanlayıcıların branş ve deneyimlerine göre dağılımı

Branş Türü ve Sınıf Düzeyi	Puanlama Deneyimine Sahip	Daha Önce Puanlama Deneyimi Olmayan	Toplam
Sınıf Öğretmeni (4. Sınıf)	2	2	4
Türkçe Öğretmeni (7. Sınıf)	2	2	4
Türk Dili ve Edebiyatı Öğretmeni (9. Sınıf)	2	2	4
Toplam	6	6	12

Veri Toplama Süreci

Veriler MEB Ölçme, Değerlendirme ve Sınav Hizmetleri Genel Müdürlüğü tarafından geliştirilen ve 2016-2017 eğitim öğretim yılı Nisan ayında 4, 7 ve 9. sınıflarda öğrenim gören öğrencilere uygulanan "Yazma Becerileri Testi"nin araştırma grubu için seçilen puanlayıcılar tarafından puanlanması ile elde edilmiştir. Puanlama her öğrenci cevabının iki puanlayıcı tarafından puanlanması şeklinde gerçekleşmiştir. İlgili veriler MEB ÖDSGM'den gerekli izinler alınarak bu çalışmada kullanılmıştır.

Veri Toplama Araçları

Kullanılacak veri toplama aracı MEB Ölçme, Değerlendirme ve Sınav Hizmetleri Genel Müdürlüğü tarafından geliştirilen ve 2016-2017 eğitim öğretim yılı Nisan ayında 4, 7 ve 9. sınıflarda öğrenim gören öğrencilere uygulanan "Yazma Becerileri Testi"dir. Puanlayıcılar öğrenci cevaplarını MEB'in öz kaynakları ile geliştirilen değerlendirme yazılımını kullanarak yine MEB tarafından hazırlanan analitik ve bütünsel dereceli puanlama anahtarlarını kullanarak çevrimiçi olarak puanlamışlardır. Buradan elde edilecek veriler ile güvenirlik analizleri yapılmıştır. İlgili yazma becerileri testine yönelik uygulama ve görevlerin hazırlanma süreci şu şekildedir (MEB, 2017):

Yazılı Anlatım Becerilerinin Ölçülmesi ve Değerlendirilmesi çalışmasının pilot uygulamasında iki okurum bulunmaktadır. Birinci oturum 7, ikinci oturum ise 1 yazma görevinden oluşmaktadır. I. oturum cümle ile paragraf görevlerinden ve II. oturum metin yazma görevlerinden oluşmaktadır. Uygulama A, B ve C formlarından oluşmak üzere aynı türde fakat farklı görevlerin bulunduğu üç kitapçık şeklinde oluşturulmuştur. Bu araştırma için alınan izin doğrultusunda, analizler yalnızca A kitapçığında bulunan veriler üzerinde yapılmıştır. Uygulama, aralarında 30 dakikalık ara bulunan ve 60 dakikadan oluşan iki ayrı oturumdan oluşmuştur. Oturumlarda yer alan görev ve içerikleri şunlardır:

Birinci Oturuma ait görevler:

a) Cümle yazma görevleri

- Kelime havuzundan seçim yaparak cümle kurma.
- Cümle bağlama veya tamamlama
- Karşılıklı konuşma (diyalog) tamamlama
- Bir görselden yola çıkarak cümle oluşturma

b) Paragraf düzeyindeki görevler

- İkna edici paragraf yazma
- Betimlemeye dayalı paragraf/Tanıtıcı paragraf oluşturma
- Dilekçe veya davetiye türünde bir ürün oluşturma

İkinci Oturuma Ait Görev:

- Metin yazma görevi

Görevlerin analiz sürecindeki kodlamaları ise şu şekilde ifade edilmiştir:

- Cümle görevleri: C1, C2, C3 ce C4 (birinci, ikinci, üçüncü ve dördüncü cümle görevi olmak üzere)
- Paragraf görevleri: Paragraf görevleri cümle görevlerinden ayrı olarak alt ölçütler içerdiğinden ölçüt harfinin baş harfi olan Ö harfi eklentisi ile kodlanmıştır. Örneğin P1Ö1 ifadesi birinci paragraf görevinin birinci ölçütünü ifade ederken P1Ö2 ifadesi ise birinci paragraf görevine ait ikinci

ölçütü ifade etmektedir. Devam eden kodlamalar bu düzene göre sunulmuştur.

- Metin görevleri: Metin görevleri paragraf görevlerinde olduğu gibi alt ölçütler barındırmaktadır ve paragraf görevine benzer şekilde kodlanmıştır. M1Ö1 metin görevinin birinci ölçütünü ifade ederken M1Ö2 birinci metin görevinin ikinci ölçütünü ifade etmektedir. Devam eden ölçütlerde aynı kodlama düzeni kullanılmıştır.

Testin ve puanlama anahtarlarının hazırlanma sürecine aşağıda yer verilmiştir.

Yazma görevlerinin geliştirilmesi. Yazma görevleri geliştirilirken oluşturulması bakımından basitten karmaşığa cümle, paragraf ve metin olmak üzere üç ayrı düzeyde görev oluşturulmuştur. Bu görevlerden paragraf ve metin görevleri kendi içinde alt kriterlere sahiptir.

Cümle düzeyindeki yazma görevlerinin geliştirilmesi. Cümle düzeyindeki görevler hazırlanırken öğrencilerin farklı becerilerine odaklanılmıştır. Bu yönde öncelikle öğrencilere sunulan kelime havuzundan seçilen kelimeleri kullanarak cümle kurabilme becerilerini ölçmeye ilişkin bir görev oluşturulmuştur. Pilot uygulamanın gerçekleştirildiği tüm sınıf düzeylerinde yer alan bu görev, öğrencilerin bir bağlamla sınırlandırılması olmaksızın basit cümleler yazabilme becerilerini belirlemek için hazırlanmıştır. Bu görev ile cümle yazma becerilerinin en temel düzeyi incelenmiştir.

Cümle düzeyinde bulunan ikinci görev öğrencilerin bir görselden yola çıkarak cümle oluşturmaları şeklindedir. Bu görev de her sınıf düzeyinde kullanılmış ve öğrenciler tümüyle sınırlandırılmadan yanıtlar istenmiştir. Bunun yanında, bu görev, 4. ve 7. sınıf düzeyindeki öğrencilerin temel düzey tanımadan üst düzey yorumlamaya kadar çeşitli zihinsel süreçlerin kullanılmasına imkân verecek şekilde oluşturulmuştur. Bu durum, cümle oluşturma bakımından birbirlerinden farklı düzeylerde bulunan öğrencilerin mevcut durumlarına göre cümle oluşturabilmelerine ilişkin bir yapı oluşturmaktadır.

9. sınıf düzeyindeki görevde ise sınıf seviyesi gözetilmiş ve nispeten üst düzeyde bir zihinsel beceri olan çıkarım yapmayı gerektiren biçimde

oluşturulmuştur. Cümle oluşturmadaki çeşitlilik ile esneklik bu sınıf seviyesinde de gözetilmiştir.

Üçüncü görev, büyük bir kısmı verilen diyalog ile kesin bir bağlamın öğrencilere sunulması şeklindedir. Bu görev, öğrencilerin oluşturacağı cümlenin, diyaloğun akışı yani verilen bağlama uygun olmasını gerektirmektedir. Bu sayede öğrencinin daha büyük bağlamları kavrayıp, verilen boşlukları bağlam kapsamında tamamlayabilme becerileri gözlenmiştir. Yine bu görev, tüm sınıf seviyelerinde kullanılmıştır. Bunun sebebi, günlük yaşamdaki yazma ve konuşma süreçlerinde en çok karşılaşılan durumun bir bağlamı kavrayıp onu en uygun cümlelerle tamamlamak ve sürdürmek olmasıdır. Başka bir deyişle bir diyaloğun bağlamı, daha önce ne söylendiği ve sonra ne söylenilmesi gerektiği ile bunun en uygun şekilde nasıl ifade edilebileceğinin belirlenmesi günlük yaşamda bir hayli önemli görülmektedir. Bu görevin amacı öğrencinin ilgili durumlarda bağlama uygun cümleleri kurup kuramadığını belirlemektir.

Son görevde, cümle akıcılığının sağlanması becerisinin ölçülmesi amaçlanmıştır. Bu görev ile 7 ve 9. sınıf seviyesindeki öğrencilerin, birbirleri ile ilişkilendirilmeye uygun kısa cümlelerin birleştirilmesiyle nispeten daha uzun fakat akıcı cümleler oluşturabilme becerilerinin belirlenmesi hedeflenmiştir. Bu görev ile sunulan cümleler birkaç kelimeden oluşan kısa cümlelerdir ve birleştirildikleri durumda bile cümlenin anlaşılmasını zorlaştıracak biçimde uzun cümleler ortaya çıkmamaktadır. 4. sınıf seviyesinde ise bu görev bulunmamakta ve yerine cümlelerin tamamlanması görevlerine yer verilmiştir. Çalışma henüz pilot uygulama düzeyinde olduğu ve nihai uygulamaya geçilmediği için örnek görev yayınlanma izni alınamamıştır.

Paragraf düzeyindeki yazma görevlerinin geliştirilmesi. Bu düzeydeki görevlerde öğrencilerin yazılı anlatımlarını etkileyen farklı becerileri ve günlük yaşamda ihtiyaç duyulabilen yazma örneklerinde görülen durumlarının tespit edilmesi öneclikli amaçtır. Bu doğrultuda bütün sınıf düzeylerinde ikna ve betimleme paragrafı yazma görevlerine yer verilmiştir. Uzunluk itibarıyla paragraf görevlerine yakın olduğu için dilekçe yazma görevi de uygulama sırasında paragraf görevlerinden hemen sonra yer almıştır. Diğerlerinden farklı olarak 4. sınıf düzeyinde davetiye, 9. sınıf düzeyinde de açıklama paragrafı yazma görevleri de yer almıştır.

İkna paragraflarında bireylerin görüşlerini çeşitli gerekçelerle destekleyerek sunmaları istendiğinden öğrencilerin paragrafı iyi yapılandırması gerekmektedir. Bu sebeple öğrencilerin hem sık kullandıkları hem de organizasyon becerilerini ortaya çıkarabilecek bir tür olan ikna tercih edilmiş ve bütün sınıf düzeylerinde kullanılmıştır.

Betimleme ise öğrencilerin ilkokul yıllarından itibaren aşina oldukları bir türdür. İlkokulda bir nesneyi, varlığı, kişiyi tanıtma şeklinde başlayan, ilerleyen yıllarda detaylı betimlemeler yapmayı gerektiren bu yazılar, farklı yazılı anlatım türlerinde de sıkça kullanılmaktadır. Üstelik betimlemelere hem öyküleyici hem de bilgilendirici metinlerde yer verilebilmektedir. Dolayısıyla yazılı anlatımda betimlemelerin önemli bir yeri olduğundan ve öğrencilerin yazılı anlatımlarında bu türe sıklıkla yer vermelerinden dolayı uygulamada betimleme görevine yer verilmiştir. Her sınıf düzeyi için o düzeyin gelişimsel özelliklerine uygun görevler hazırlanmıştır.4. sınıf düzeyinde tanıtma yazıları şeklinde tasarlanan görevler 7 ve 9. sınıf düzeylerinde betimleme paragrafı yazmayı gerektirecek şekilde hazırlanmıştır.

Günlük yaşamda insanlar birçok alanda davetiye ile karşılaşmaktadır. Bu bakımdan kullanılma durumu dikkate alınarak bu türün 4. sınıf düzeyindeki görevlerden biri olması uygun görülmüştür.

9. sınıfta kullanılan açıklama paragrafının amacı öğrencilerin bir konudaki duygularının veya düşüncelerinin kısa ve öz olarak planlanıp bir paragraf hâlinde ifade edilme durumunu tespit etmektir. Bu tür kısa ve yoğun anlatımların günlük yaşamda kullanılma durumları giderek arttığı için açıklama paragrafı lise düzeyindeki görevlerden biri olarak belirlenmiştir.

Metin düzeyindeki yazma görevlerinin geliştirilmesi. Metin düzeyindeki görevler okul ve iş yaşamındaki yazma durumlarından hareketle belirlenmiştir. Bu kapsamda öğrencilerin ilkokuldan itibaren en sık yazdıkları tür olan hikâye yazma görevine bütün sınıf düzeylerinde yer verilmiştir. Ayrıca ilkokul 4. sınıf için hazırlanan metin yazma görevlerinin tamamı hikâye türünde metin yazmaya yöneliktir.

7. sınıf düzeyinde de öğrencilerin en rahat yazabilecekleri türlerden biri olan anı türünde bir metin yazma görevi kullanılmıştır.

Uygulamada son olarak hem 7. hem de 9. sınıflarda birer bilgilendirici metin yazma görevine yer verilmiştir. Her iki sınıf düzeyinde de bilgi vermeye yönelik metinlerin Türkçe ve Türk Dili ve Edebiyatı derslerindeki okuma ve yazma çalışmalarında sıkça kullanılmaya başlanmasından dolayı öğrencilerin bu tür metinleri yazma konusundaki becerilerini tespit etmek amaçlanmıştır.

Görevlerin puanlanmasında cümle, paragraf ve metin görevleri için ayrı olmak üzere üç farklı puanlama anahtarı kullanılmıştır. İlgili puanlama anahtarları analitik ve bütünsel olarak görevin niteliğine göre belirlenmiştir.

Cümle görevlerinin puanlanması Yazılı anlatım becerisinin en temel boyutlarından birisi cümle yazımıdır. Cümle boyutuna ilişkin görevler paragraf ve metin görevlerine kıyasla daha az bileşen içerir. Öğrenci yanıtı bir ya da iki sözcük ile cümleden oluşabilmektedir. Bu tip görevlerde öğrenci performansını alt bileşenlerine ayırmadan bir bütün olarak değerlendirmek daha doğru olduğu için cümle görevlerinde bütünsel puanlama anahtarının kullanımının yerinde olduğu düşünülmüştür.

Paragraf görevlerinin puanlanması. Paragraf görevleri yapısı gereği cümle görevlerinden daha karmaşıktır ve içerisinde ölçülmesi önemli olan başka bileşenleri de içerebilmektedir. Bu nedenle paragraf görevlerinde analitik DPA'lar kullanılmıştır.

Metin görevlerinin puanlanması. Metin yazımı diğer yazılı anlatım becerileri içerisinde en kapsamlı olanıdır ve pek çok bilgi ve becerinin bir arada kullanılmasını gerektirir. Bu yapısı ile metin görevlerinin puanlanmasında analitik DPA'ların kullanılması daha uygundur.

Verilerin Analizi

Öğretmenlerin yazma görevlerini puanlamaları sonucu elde edilen veriler incelenmiş ve kayıp veri olmadığı için bu yönde bir ayıklama veya düzeltme yapılmamıştır. İlgili verilerin analizi için kullanılan üç farklı modele yönelik analiz süreçleri aşağıda ifade edilmiştir.

Cohen'in ağırlıklandırılmış Kappa katsayısına (κ) yönelik kestirimde bulunmak için bir çalışma tasarlarken, örneklem boyutu, κ 'nın standart hatası önceden belirlenmiş bir değeri aşmayacak şekilde seçilmelidir (Cantor, 1996). κ için minimum örneklem sayısı belirlenirken, yapılacak olan araştırmanın niteliği büyük

önem taşımaktadır. Örneğin, klinik alanlarda insan hayatını doğrudan etkileyebilecek durumlarda uyumun yüksek olması istenen bir durumdur. Bu araştırmada minimum κ değeri olarak 0,55, beklenen κ değeri 0,8, manidarlık seviyesi olarak α = 0,05 (one tailed) ve %80 güçlükte bir çalışma yürütüldüğü varsayılırsa, Arifin (2018) tarafından önerilen web tabanlı örneklem hesaplama uygulamasında 88 kişilik bir örneklemin çalışma için uygun olacağı çıkarımı yapılabilir. Her sınıf düzeyinde 80 öğrenci bulunduğu göz önüne alınırsa örneklem büyüklüğünün kabul edilebilir olduğu varsayılabilir.

Alan yazın incelendiğinde, Krippendorff'un Alpha katsayısı için örneklem büyüklüğünün küçük veya büyük olabileceği ve minimum bir değer gerektirmediği belirtilmiştir (Krippendorff, 2011). Krippendorff'un Alpha katsayısı için veriler sınıflama, sıralama, eşit aralık veya oranlı düzeyde olabilir. Araştırma verileri sıralama düzeyinde olduğu için analize uygun bulunmıştur. Analiz SPSS v26 paket programında Krippendorff Alpha istatistiği için özel olarak yazılmış bir syntax kullanılarak gerçekleştirilmiştir.

Son olarak, bir G çalışması yürütebilmek için gerekli olan örneklem sayısı hakkında Webb, Rowley ve Shavelson (1988) tarafından ileri sürülen bir öneri, G çalışmalarının her yüzey için en az 20 kişi ve 2 koşulu içermesidir. Bu nedenle tek yüzeyli bir tasarım için toplam 40 verinin yeterli olabileceğini ifade etmişlerdir. Bununla birlikte Lei, Smith ve Suen (2007), G kuramının tek bir kişi için elde edilen zaman serisi verilerine uygulanmasının, durumlar yeterli şekilde örneklendiğinde ve yüzey sabit olarak ele alındığında bile uygun olabileceğini öne sürmüşlerdir. Atılgan (2013) tarafından yapılan ve G ve Phi katsayılarının örneklem büyüklüklerinden etkilenme düzeyine bakıldığı bir çalışma ile ortaya konmuştur. Bu çalışmada, n=30, 50, 100, 200, 300, 400 gibi örneklem sayıları ile denemeler yapılmış ve n=30 kişilik bir örneklem büyüklüğünün G ve Phi katsayılarını kestirmede yetersiz olduğu ve n=50, 100 200 ve 300 gibi örneklem büyüklüklerinde katsayıların yansız olarak kestirilmesinin uygun olduğu sonucuna varılmıştır. Dolayısıyla her sınıf düzeyinde 80 kişilik bir örneklem büyüklüğüne sahip olan bu çalışmanın nispeten uygun değerlere sahip olduğu söylenebilir.

Bölüm 4

Bulgular ve Yorumlar

Bu bölüm altında araştırma problemine yönelik analizler, bulgular ve bunlara yönelik yorumlar bulunmaktadır.

Alt Problem 1' e Yönelik Bulgu ve Yorumlar

"Alt problem 1. Sınıf öğretmenlerinin puanlayıcılar arası uyum düzeyi nedir?" Sınıf öğretmenlerinin puanlayıcılar arası uyum düzeyleri analizi bulguları; Cohen'in Ağırlıklandırılmış kappa, Krippendorff'un Alpha ve G kuramına göre, deneyim durumları göz önüne alınarak ayrı ayrı verilmiştir.

Puanlama deneyimine sahip ve daha önce puanlama deneyimi olmayan sınıf öğretmenlerinin puanlayıcılar arası uyum düzeyleri. Cohen'in Ağırlıklandırılmış kappa, Krippendorff'un Alpha ve G kuramı istatistik düzeyleri SPSS v26 paket programı ve EduG yazılımı kullanılarak kestirilmiştir. Yazma Becerileri Testinde bulunan cümle görevleri birer birer, paragraf görevleri ve metin görevi analitik puanlama anahtarı kullanılarak puanlandığı için her bir kriter birer madde şeklinde ele alınmış olup, bu puanlamalara yönelik analizler Tablo 7'de verilmiştir. G kuramına göre yapılacak analizde EduG yazılımının bütün maddeleri tek seferde değerlendirmesi ve madde bazında bir varyans kaynağı vermemesi sebebiyle ayrı bir tablo hazırlanmış ve yorumlanmıştır. Krippendorff'un Alpha katsayısı hesaplanırken, sonuçların genelleneceği evreni temsil eden "bootstrap sample" düzeyi en yüksek seçenek olan 10.000 olarak alınmıştır. Alınan bootstrap sample sayısı ne kadar büyükse, çıkarımsal istatistikler o kadar doğru olur. 10,000 ve üzeri bootstrap sample değeri çok minimal değişikliklerle sonuçlanacağı için bu sayı yeterli olarak kabul edilmektedir (Krippendorff ve Hayes, 2007).

Tablo 7

Sınıf Öğretmenlerinin Puanlayıcılar Arası Uyum Düzeyleri

Görev		Deneyimsiz				Deneyimli			
		κ	κ (p)	α	α (q)	κ	κ (P)	α	α (q)
C1		0,718	0,000	0,805	0,003	0,593	0,000	0,682	0,399
C2		0,619	0,000	0,622	0,667	0,519	0,000	0,524	0,976
C3		0,694	0,000	0,677	0,437	0,501	0,000	0,611	0,781
C4		0,137	0,125	0,187	0,999	0,364	0,000	0,472	0,997
P1	Ö1	0,574	0,000	0,643	0,597	0,436	0,000	0,523	0,947
	Ö2	0,526	0,000	0,639	0,615	0,423	0,000	0,511	0,951
	Ö3	0,490	0,000	0,616	0,704	0,486	0,000	0,537	0,928
	Ö4	0,497	0,000	0,608	0,715	0,447	0,000	0,530	0,931
P2	Ö1	0,314	0,000	0,420	0,978	0,456	0,000	0,391	0,997
	Ö2	0,381	0,000	0,449	0,967	0,439	0,000	0,626	0,712
	Ö3	0,367	0,000	0,462	0,966	0,435	0,000	,05907	0,819
	Ö4	0,346	0,000	0,410	0,978	0,504	0,000	0,621	0,725
P3	Ö1	0,429	0,000	0,556	0,952	0,192	0,021	0,211	1,0
	Ö2	0,540	0,000	0,570	0,870	0,242	0,001	0,264	1,0
	Ö3	0,541	0,000	0,615	0,753	0,189	0,019	0,231	1,0
	Ö4	0,481	0,000	0,638	0,694	0,160	0,010	0,215	1,0
M1	Ö1	0,432	0,000	0,361	0,982	0,392	0,000	0,468	0,975
	Ö2	0,567	0,000	0,578	0,784	0,395	0,000	0,460	0,974
	Ö3	0,337	0,000	0,409	0,996	0,105	0,015	-0,200	1,0
	Ö4	0,407	0,000	0,443	0,988	0,188	0,003	0,138	1,0
	Ö5	0,446	0,000	0,465	0,961	0,207	0,001	0,210	1,0
	Ö6	0,362	0,000	0,417	0,992	0,144	0,014	0,054	1,0
	Ö7	0,243	0,000	0,308	1,0	0,118	0,062	0,042	1,0
	Ö8	0,456	0,000	0,515	0,950	0,115	0,043	-0,008	1,0

Alt Problem 1-a, Cohen'in ağırlıklandırılmış kappa istatistiğine göre puanlama deneyimine sahip ve daha önce puanlama deneyimi olmayan sınıf öğretmenlerinin puanlayıcılar arası uyum düzeyi nedir? Tablo 7'de Yazma Becerileri Testini alan 4. sınıf öğrencilerinin verdiği yanıtlar doğrultusunda deneyimli ve deneyimsiz puanlayıcıların verdikleri puanların birbiriyle uyumunu göstermek amacıyla Cohen'in kappa (κ) ve Krippendorff'un Alpha (α) katsayılarına ait analiz sonuçları yer almaktadır. Tablo 7 incelendiğinde κ'ya göre deneyimsiz puanlayıcılarda en fazla uyumun gözlendiği madde κ = 0,718 ile C1 maddesidir. Deneyimli puanlayıcılar arasında da C1 maddesi κ = 0,593 ile en çok uyumun gözlendiği madde olmuştur. C1 maddesinde deneyimsiz puanlayıcıların uyumlarının iyi olduğu görülmektedir. Buna karşın, deneyimli puanlayıcılar arasındaki uyum yeterli düzeyde kalmıştır. Deneyimsiz puanlayıcılar arasında C1 maddesini sırasıyla C3 (0,694), C2 (0,619), P1Ö1 (0,574) ve M1Ö2 (0,567) madde ve ölçütleri takip etmektedir. C2 ve C3 maddeleri iyi uyum düzeyine sahipken P1Ö1 ve M1Ö2 ölçütleri yüksek olmasa da yeterli olarak kabul edilmektedir. Deneyimli puanlayıcılar arasında C1 maddesinden sonra en uyumlu olunan maddeler sırasıyla C2 (0,519), P2Ö4 (0,504), C3 (0,501) ve P1Ö3 (0,486) madde ve ölçütleridir. Bu dört madde de yeterli uyum düzeyine sahip olarak yorumlanabilir.

Deneyimsiz puanlayıcılar arasında κ istatistiğine göre en düşük uyumun gözlendiği madde 0,137 ile C4 maddesidir. Deneyimli puanlayıcıların bu maddedeki uyumu nispeten yüksek olmakla birlikte uyum gücü açısından zayıf kalmıştır. Deneyimli puanlayıcılar arasında en düşük uyum M1Ö3(0,105) ölçütünde gözlenmiştir. Bu değerler uyum açısından kabul edilebilir düzeyin çok altındadır ve madde, puanlama anahtarı veya puanlama yönergesi ile ilgili ciddi bir sıkıntı olduğuna işaret etmektedir. Deneyimsiz puanlayıcılarda C4 maddesini sırasıyla M1Ö7 (0,243), P2Ö1 (0,314), M1Ö3 (0,337) ve P2Ö4 (0,346) ölçütleri takip etmektedir. Deneyimli puanlayıcılarda en düşük uyumun gözlendiği madde olan M1Ö3 ölçütünü sırasıyla M1Ö8 (0,115), M1Ö7 (0,115), M1Ö6 (0,144) ve M1Ö4(0,188) ölçütleri takip etmiştir. Bu ölçütlerdeki uyum değeri de κ istatistiğine göre kabul edilebilir düzey olan 0,40 seviyesinin altındadır. Genel olarak bakıldığında öğretmenlerin, 4. cümle görevi hariç olmak kaydıyla, bütünsel DPA kullanılan cümle görevlerinde analitik DPA kullanılan paragraf ve metin görevlerine oranla daha yüksek bir uyum içerisinde oldukları söylenebilir. Deneyimsiz

öğretmenlerin puanladığı 2. Paragraf görevindeki ölçütlerin tümünde düşük uyum gözlenmiştir. Metin görevindeki uyum katsayıları ise nispeten yeterli olarak değerlendirilebilir. Deneyimli öğretmenlerde ise metin görevinin ölçütlerinde son derece düşük bir uyum gözlenmiştir. Bunun sebebi yönergelerle ilgili hatalar olabileceği gibi geçmiş puanlama deneyimlerine dayanarak yapılan puanlama anahtarının yanlış yorumlanması da olabilir. Metin görevinin yanında genel olarak deneyimli öğretmenlerin uyum düzeyleri, puanlama deneyimi olmayan öğretmenlere oranla, Cohen'in Kappa istatistiğine göre oldukça zayıf kalmıştır. Düşük kappa değerleri, araştırılan ölçme durumunun, örneklemde bulunan bireyler arasında net ayrımlar yapamadığını gösterir ki bu ayrımı gerçekleştirmek oldukça zor ve nadirdir. Ek olarak, puanlayıcıların birbirine yakın olan kategorileri ayırt etmekte zorlandığı şeklinde de yorumlanabilir (Kreamer ve Noda, 2002; Vach, 2005; Darroch ve McCloud, 1986).

Alt problem 1-b, Krippendorff'un Alpha katsayısı istatistiğine göre puanlama deneyimine sahip ve daha önce puanlama deneyimi olmayan sınıf öğretmenlerinin puanlayıcılar arası uyum düzeyi nedir? Tablo 7'deki maddelere ait Krippendorff α katsayıları incelendiğinde deneyimsiz puanlayıcılar arasında en yüksek uyumun gözlendiği madde 0,805 ile C1 maddesidir. Bu maddeyi sırasıyla C3 (0,677), P1Ö1 (0,643), P1Ö2 (0,639), P3Ö4 (0,638) ve P1Ö3 (0,616) madde ve ölçtüleri takip etmektedir. Krippendorff (1995)'a göre 0,67 değerinin altındaki uyum zayıf olarak değerlendirilmektedir. Bu bağlamda, C1 maddesi yüksek ve P1Ö1 ölçütü orta düzeyde uyum düzeyine sahiptir. Bunların dışında kalan ölçüt ve maddelerdeki puanlayıcı uyumları genel olarak zayıf kalmıştır. En zayıf uyumun olduğu madde κ'da olduğu gibi 0,187 değeriyle C4 maddesidir. Bu maddeyi sırasıyla M1Ö7 (0,308), M1Ö1 (0,361), M1Ö3 (0,409) ve P2Ö4 (0,410) takip etmiştir. Genel olarak bakıldığında en düşük uyum düzeylerinin metin ve 2. Paragraf görevlerine ait ölçütlerde olduğu söylenebilir. Krippendorff α katsayısı incelenirken dikkate alınması gereken ikinci bir olasılık (q) indeksi mevcuttur. Bu indeks, tüm popülasyonun test edilmesi durumunda, uyumun α istenen en düşük düzey olan 0,67 değerinin altında olma ihtimalini temsil etmektedir. İlgili q değerlerine incelendiğinde, C1 maddesinde bu olasılık değerinin 0,003 olduğu görülebilir. Bu, tüm popülasyonun test edilmesi durumunda, uyum düzeyinin 0,67'nin altında olma ihtimalinin 3‰ (binde 3) olduğu şeklinde yorumlanır. Bu maddeyi takip eden P1Ö1

ölçütünde bu değer 0,597 olarak bulunmuştur. Tüm popülasyonun test edilmesi durumunda, P1Ö1 ölçütündeki uyum düzeyi 59% olasılıkla 0,67 değerinin altında olacaktır. Nispeten en kötü çalışan C4 maddesi 0,99 q değeriyle tüm popülasyonun test edilmesi durumunda 99% olasılıkla 0,64 değerinin altında olacaktır. M1Ö7 ölçütündeki uyum düzeyi tüm popülasyonun test edilmesi durumunda %100 olasılıkla 0,67 değerinin altında olacaktır.

Tablo7'de deneyimli puanlayıcılara ait Krippendorff α katsayıları incelendiğinde, bu puanlayıcılar arasında en yüksek uyumun gözlendiği madde 0,682 ile C1 maddesidir. Bu maddeyi sırasıyla P2Ö2 (0,626), P2Ö4 (0,621), P2Ö3 (0,590), P1Ö3 (0,537) ve P1Ö4 (0,530) ölçütleri takip etmektedir. Krippendorff (1995)'a göre 0,67 değerinin altındaki uyum durumunun zayıf olduğu düşünülürse, 1.cümle görevi orta düzeyde uyuma sahipken diğer tüm ölçüt ve maddelerdeki puanlayıcı uyumları zayıf olarak değerlendirilebilir. En zayıf uyumun olduğu ve negatif değer almış olan M1Ö3 (-0,200) ve M1Ö8 (-0,008) ölçütleridir. Negatif Alpha değerleri şans uyumunun altında bir uzlaşmaya işaret etmektedir. Alpha, puanlayıcılar sürekli olarak farklı puanlar verdiklerinde negatif değerler alabilir (Krippendorff, 2008). Bu durumda bu iki ölçüt için sürekli bir uyumsuzluktan bahsedilebilir. Bu ölçütleri sırasıyla M1Ö7 (0,042), M1Ö6 (0,054), M1Ö3 (0,105) ve M1Ö6 (0,144) takip etmiştir. En düşük uyum düzeyleri metin ve 3. paragraf görevlerine ait ölçütlerde görülmektedir. Olasılık (q) indeksleri incelendiğinde, 3. Paragraf görevine ait tüm ölçütlerde, tüm popülasyonda yapılacak bir test uygulamasında uyum düzeyinin 0,67 olan sınır değerinin altında olma ihtimali %100 olarak bulunmuştur. Bununla birlikte, 1. Cümle görevi hariç neredeyse tüm maddelerde uyum düzeyinin 0,67 altında çıkması olasılığı beklenmektedir.

Sınıf öğretmenlerine yönelik yapılan uyum düzeyleri karşılaştırması sonucunda genel olarak deneyimsiz puanlayıcılar deneyimli puanlayıcılara göre daha uyumlu puanlama gerçekleştirmiştir. Deneyimsiz öğretmenlerin uyumsuzluk düzeyleri en çok 2. Paragraf görevi ölçütlerinde iken, deneyimli öğretmenler en yüksek uyumsuzluğu metin görevine ait ölçütlerde göstermiştir.

Alt Problem 1-c, G kuramına göre puanlama deneyimine sahip ve daha önce puanlama deneyimi olmayan sınıf öğretmenlerinin G ve phi katsayısı düzeyleri nedir? Birinci alt probleme ait son soru olan G kuramına yönelik analizler

gerçekleştirilmiştir. Bu analiz sonuçlarının sunulmasından önce oluşturulan desen ve yüzeylere ilişkin bilgiler Tablo 8'de sunulmuştur.

Tablo 8

Birinci Alt Probleme Ait G Kuramı Deseninde Bulunan Yüzeyler

Yüzey	Etiket	Düzey	Evren
Birey	B	80	SONSUZ
Madde	M	24	SONSUZ
Puanlayıcı	P	2	SONSUZ

Çalışmada yüzey (facet) olarak bireyler, görevler (madde) ve bu çalışmanın başlıca konusu olan puanlayıcılar belirlenmiştir. Bu yüzeylere ait düzeyler (level) ise öğrenci sayısı olan 80 ve puanlayıcı sayısı olan 2'dir. Burada dikkat edilmesi gereken bir nokta madde sayısı olarak 24'ün belirlenmiş olmasıdır. Hatırlanacağı gibi asıl uygulamada 4 cümle, 3 paragraf ve 1 metin görevi bulunmakta ve cümle görevleri bütünsel DPA ile puanlanırken paragraf ve metin görevleri analitik DPA'lar ile puanlanmaktadır. Analitik DPA'lar paragraf görevleri için 4, metin görevleri için 8 ayrı ölçütten meydana gelmektedir. Söz konusu bütünsel DPA'lar tek bir puandan oluşurken analitik DPA'larda bulunan her ölçüt için puanlayıcılar ayrı ayrı puanlama yapmaktadır. Dolayısıyla her ölçütte karşılaşılan uyum düzeyinin dikkate alınması için her ölçüt bir madde gibi düşünülmüş ve bu sebeple 24 düzeyden oluşan bir yüzey oluşturulmuştur (Cümle görevi için 4, her bir paragraf görevi için 4 olmak üzere toplamda 12 ölçüt ve metin görevi için 8 ölçüt şeklinde). Genellenmek istenen evren tüm öğrenciler, aynı özelliği ölçen maddeler ve farklı puanlayıcılar olduğu için evren sonsuz seçilmiştir. Eğitimde ölçme objesi birey olduğu için ölçme objesi olarak bireyler alınmıştır. Tüm bireyler (80 öğrenci), puanlamaya katılan tüm puanlayıcılar (2 öğretmen) tarafından her madde ve ölçüt (4 madde ve 20 ölçüt olmak üzere toplam 24) için puanlandığından desen tümüyle çaprazlanmış bxmxp ve karma desen olarak belirlenmiştir. Son olarak, ölçme desenindeki puanlayıcılar, maddeler ve bireyler evrenden tesadüfi olarak alınabilir ve değiştirilebilir bir durumda olduğu için ve bulgular uygulanan durumun ötesine genellenmek istendiği için, yüzeyler tesadüfi (random) olarak ele alınmış ve her sınıf düzeyi ve deneyim grubu için ayrı G çalışmaları analizi yapılmıştır. Tüm analizler EduG (Swiss Society for Research in Education Working Group, 2006) yazılımı ile gerçekleştirilmiştir.

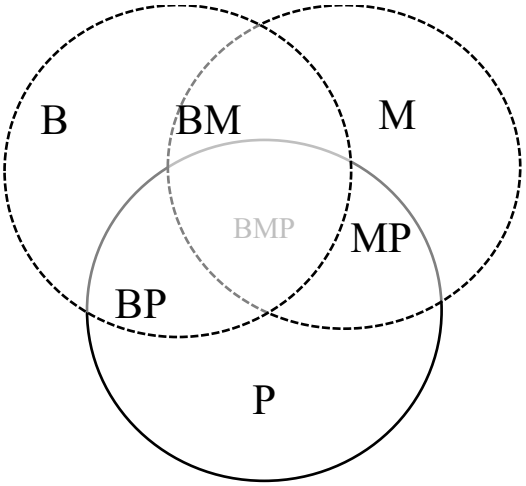
G çalışması için tahmini varyans bileşenleri; Birey, madde, puanlayıcı, birey x madde, birey x puanlayıcı, puanlayıcı x madde ve kalan üç yönlü birey x puanlayıcı x madde ve hata etkileşiminin toplamından oluşan yüzeyler olmak üzere hesaplanmıştır (bkz. Tablo 9).

Tablo 9

b x m x p Tümüyle Çaprazlanmış Karma Model Desenine ait Varyans Kaynakları

Yüzey	Varyansın Türü
Birey (*b*)	Ölçme Objesi
Madde (*m*)	Madde zorluğuna bağlı tutarsızlıktan kaynaklanan varyans
Puanlayıcı (*p*)	Puanlayıcılar arasındaki tutarsızlıktan kaynaklanan varyans
b x m	b x m etkileşiminden kaynaklanan varyans
b x p	b x p etkileşiminden kaynaklanan varyans
m x p	m x p etkileşiminden kaynaklanan varyans
b x m x p + e	b x m x p etklişeminden kaynaklanan varyans ve hata toplamı

D (Karar) çalışmaları. G çalışmalarından elde edilen sonuçları kullanarak, her deneyim düzeyi için bir dizi D (karar) çalışması yapılmıştır. Hatırlanacak olursa, D çalışmasının amacı, belirli bir amaç için genelleme evrenini tanımlamak ve daha sonra, ölçümün çeşitli yüzeylerinin değiştirilmesinin, elde edilen verilerden alınan kararları nasıl etkilediğini analiz etmektir. Puanlayıcı yüzeyi, kabul edilebilir düzey olan .80 veya üzeri genellenebilirlik (bağıl kararlar için G ($E\rho^2$)) ve güvenirlik (mutlak kararlar için Phi (Φ)) katsayıları (Ulrich, Ulrich ve Branta, 1988) elde etmek için optimizasyon seçeneği ile değiştirilmiştir. Bu sayede iki puanlayıcının mevcut ölçme durumu için güvenilir sonuçlar üretip üretemeyeceğini ve şayet üretemeyecekse bu sonuca ulaşmak için kaç puanlayıcıya ihtiyaç duyulacağını görmek amaçlanmıştır.



Şekil 4. Varyans dağılım venn diyagramı.

G kuramı altında yapılan analizlere yönelik tümüyle çaprazlanmış bxmxp desenine ilişkin varyans bileşenleri bulguları Tablo 10'da gösterilmiştir.

Tablo 10

Sınıf Öğretmenlerinin Tümüyle Çaprazlanmış bxmxp Desenine Ait Kestirilen Varyans Bileşenleri

Varyans Kaynağı	Varyans	% Varyans	Varyans	% Varyans
B	0,19172	10,4	0,33722	18,5
M	-0,00214	0,0	0,42548	23,4
P	-0,00056	0,0	-0,00052	0,0
BM	1,02764	55,7	0,28342	15,6
BP	0,03670	2,0	0,09708	5,3
MP	0,00389	0,2	0,07335	4,0
BMP	0,58496	31,7	0,60379	33,2
Toplam		%100		%100
$E\rho^2$	0,72		0,82	
Φ	0,72		0,79	

Tablo 10'da deneyimsiz puanlayıcılara ait sütun incelenirse, bireyler için kestirilen varyans, toplam varyans içinde %10,4'lük bir değere sahip olduğu görülebilir. Bu varyans, evren puan varyansı olarak da adlandırılmaktadır. KTK'da gerçek puan varyansını ifade eden evren puanı varyansı bireylerin ölçülen özellik bakımından ne ölçüde farklılaştıklarının bir derecesidir. Ölçme işlemi sonucunda bireylerin ölçülmek istenen özellikleri arasındaki farklılığın ortaya konulabilmesi beklenir (Atılgan, 2019). Bu varyans değerinin toplam varyansın içindeki göreceli değerinin büyüklüğü, bireylerin sistematik farklılıklarının ortaya konulabildiğini ve gözlenen puanın evren puanını temsil etme gücünü gösterir. Bu varyans değerinin toplam varyans içindeki değerinin yüksek olması istenen bir durumdur. Tablo 10 incelendiğinde, gözlenen puanların gerçek puanı temsil etmede yetersiz kaldığı söylenebilir. Bununla birlikte madde (-0,00214) ve puanlayıcılardan (-0,00056) kaynaklanan varyans bileşenlerinin negatif değerler aldığı görülmektedir. Bir varyans bileşenin tahmininin, grup-içi ortalamaların karesinin gruplar-arası ortalama karelerinden çıkararak kestirildiği düşünüldüğünde, küçük negatif değerler bulmak olasıdır. Diğer bir deyişle, bazı ölçme durumlarında, grup-içi farklılıklar gruplar-arası farklılıklardan daha büyük olabilir. Çok küçük negatif varyans bileşenlerinin gözlenmesi önemsiz olarak görülebilmektedir (Briesch, Swaminathan, Welsh ve Chafouleas, 2014). Bununla birlikte, büyük negatif değerlerin genellikle örnekleme hatasının varlığına ya da modelin yanlış tanımlandığına işaret edebileceği unutulmamalıdır (Shavelson ve Webb, 1991).

Görevlerden kaynaklanan varyans düzeyi, seçkisiz olarak belirlenen herhangi bir görev ortalamasının evrendeki tüm maddelerin ortalaması şeklinde beklenen ortalamadan farkına işaret eder. Bu varyansın büyük olması görevlerin güçlük düzeylerinin arasındaki farklılıkların artmasını, sıfıra yaklaşması durumunda görev güçlüklerinin birbirine benzer olduğunu temsil eder (Briesch ve arkadaşları, 2014).. Madde kaynaklı varyans oranı %0,0 olarak bulunmakla birlikte bu durum madde güçlükleri arasında herhangi bir farklılık olmadığı anlamına gelmektedir.

Puanlayıcılardan kaynaklanan varyans yüzdesi de görevlerde olduğu gibi %0 bulunmuştur. Bu varyans puanlayıcıların tüm bireyler arasında verdikleri puanların diğer puanlayıcılara göre hoşgörülü veya katı olup olmadıklarını ifade eder. Bu varyansın toplam varyans değeri içinde düşük çıkması puanlayıcılar arasındaki değişimin fazla olmadığını, büyük çıkması ise bazı puanlayıcıların diğerlerine göre

daha hoşgörülü veya katı puanladığı şeklinde yorumlanır (Brennan, 2001). %0'lık bir varyans yüzdesi puanlayıcıların diğer puanlayıcılara göre katı veya hoşgörülü olma açısından büyük farklılık göstermediği şeklinde yorumlanabilir.

Birey ve görev ortak etkisinin (BM) %55,7 ile en yüksek varyans yüzdesine sahip olduğu görülmektedir. Bu yüzdenin toplam varyans içindeki payının büyük olması bireylerin bir görevden diğerine bağıl durumlarındaki farklılığın yükseldiği anlamına gelir. Bu, bireyin bir görevi kolay bir şekilde yanıtlarken başka bir görevde zorlandığını göstermektedir.

Birey-puanlayıcı (BP) ortak etkisi için bulunan varyans bileşeni %2'dir. Bu varyans bileşeninin toplam varyans içindeki değerinin az olması belirli puanlayıcıların belirli bireylere yönelik puanlamalarının diğer puanlayıcılara göre değişiklik göstermediği şeklinde yorumlanabilir.

Görev-puanlayıcı (MP) ortak etkisine yönelik bulunan varyans yüzdesi %0,2'dir. Bu varyans yüzdesinin toplam varyans yüzdesi içindeki değerinin az olması puanlayıcıların görevleri puanlarken bir görevden diğerine benzer şekilde ve kararlı puanlamalar yaptıkları şeklinde yorumlanabilir.

Varyans bileşenlerindeki sonuncu varyans kaynağı olan ve birey, görev ve puanlayıcı ortak etkisine atfedilen BMP varyansı ölçülemeyen varyans kaynakları olarak ifade edilir. Ölçülemeyen varyansa yönelik kaynaklar sistematik veya sistematik olmayan (tesadüfi) sebeplerden kaynaklanıyor olabilir (Brennan, 2001). Tablo 10 incelendiğinde ölçülemeyen varyans kaynakları toplam varyans içinde %31,7'lik bir paya sahiptir. Bu, ölçme durumundaki ölçülemeyen değişkenlik kaynaklarının nispeten yüksek olduğunu göstermektedir.

Yapılan G çalışması sonucu ölçme objesine bağlı varyans bileşenlerinin toplamı sonucu elde edilen bağıl hata varyansı 0,07336 ve G ve phi katsayıları 0,72 olarak kestirilmiştir. Genel olarak incelendiğinde, deneyimsiz sınıf öğretmenleri, Yazma Becerileri Testinde aşırı katı veya hoşgörülü olmadan ve bireyler ile görevler arasında değişiklik göstermeden puanlama gerçekleştirmişlerdir

Tablo 10'daki deneyimli puanlayıcılara ait sütun incelendiğinde, toplam varyansın %9,3'ü puanlayıcılardan kaynaklanan etkilere atfedilmiştir. Özel olarak incelenirse, birey ve puanlayıcı (%5,3) ve madde ve puanlayıcı (%4,0) ortak etkilerine yönelik varyans kaynakları, puanlayıcı (%0) kaynaklı varyans ile birlikte

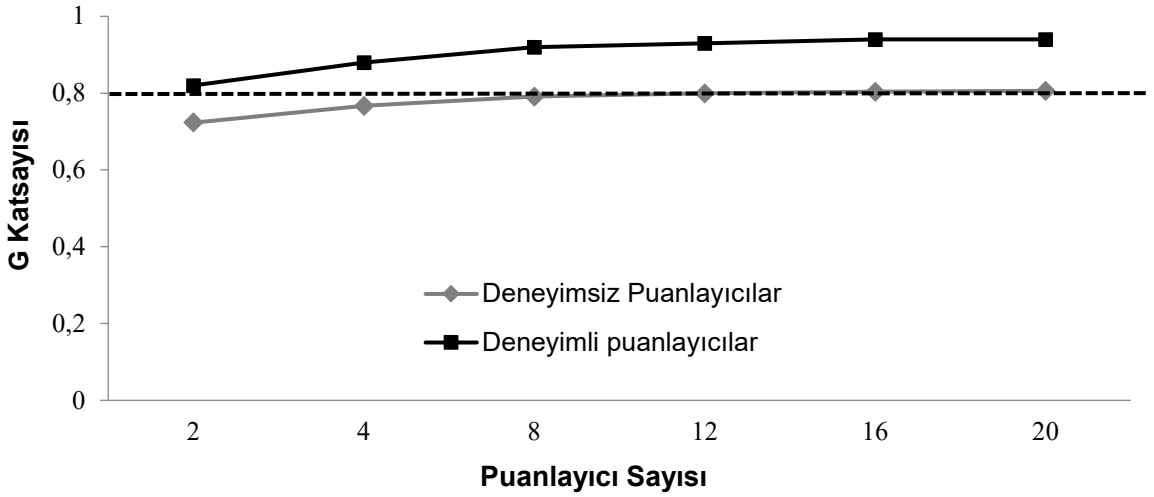
toplam varyansa en az katkı sağlayan etki olarak görülebilir. Bu sonuca dayanarak, öğretmenlerin bireyleri ve görevleri sıralamada büyük farklılıklar göstermedikleri söylenebilir. Buna karşın, en büyük varyans yüzdesinin birey, madde ve puanlayıcı (BMP) ortak etkisine (%33,2) bağlı olduğu görülmektedir. Bu, ölçülemeyen değişkenlik kaynaklarının çokluğuna işaret etmektedir. Toplam varyans içindeki en düşük paya sahip olan puanlayıcı (P) yüzeyine atfedilebilen varyans yüzdesi (%0), öğretmenlerin gerçekleştirdiği puanlamaların genel olarak önemli ölçüde değişmediğini, katılık veya hoşgörü göstermediğine işaret etmektedir. Birey ve puanlayıcı (BP) arasındaki etkileşimin, gözlemlenen varyansın %5,3'ünü açıkladığı düşünüldüğünde, öğretmenlerin belirli öğrencilere yönelik puanlamalarını bir miktar değişmekle birlikte kayda değer ölçüde değişmediği görülmektedir. Madde ve puanlayıcı (MP) arasındaki etkileşim, toplam varyansa %4 katkıda bulunmuştur. Bu varyans yüzdesinin küçük olması, maddeler arasındaki puanlama davranışında küçük değişiklikler olduğunu göstermektedir. Ölçme objesi olan bireylerin toplam varyans içerisindeki oranı %18,5'tir. Bu puanlayıcıların öğrencilerin arasındaki farklılıkları nispeten ayırabildiğini gösterir. Son olarak, madde yüzeyinin (M) toplam varyansa en büyük ikinci (%23,4) derecede katkıda bulunduğu ve bu da maddelerin bir maddeden diğerine genel olarak değişiklik gösterdiğine (zorluk veya ayırt edicilik gibi) işaret etmektedir.

Deneyimli puanlayıcılara yönelik yapılan G çalışması sonucu ölçme objesine bağlı varyans bileşenlerinin toplamı sonucu elde edilen bağıl hata varyansı 0,07293 ve G katsayısı 0,82 ve Phi katsayısı 0,79 olarak kestirilmiştir. Kestirilen G katsayısının kabul edilebilir düzeyde olduğu yorumu yapılabilirken Phi katsayısının istenen 0,80 düzeyinin üzerinde olması için gereken durumların denendiği bir D çalışması yürütülmüştür. Deneyimli sınıf öğretmenleri, Yazma Becerileri Testinde, G kuramına göre, aşırı katı veya hoşgörülü olmadan ve bireyler ile görevler arasında değişiklik göstermeden puanlama gerçekleştirmişlerdir. Ayrıca, G kuramı altında yapılan analiz sonuçlarına göre, deneyimli sınıf öğretmenlerinin yaptığı puanlamalara yönelik G katsayısı (0,82) ve Phi katsayısı (0,79), deneyimsiz sınıf öğretmenlerine ait G (0,72) ve Phi (0,72) katsayılarına göre daha yüksek bulunmuştur. 0,70 ve üzerinde olan güvenirlik katsayıları uygun kabul edilmekle birlikte, arzu edilen 0,80 Phi güvenirlik düzeyine ulaşmak için D çalışmaları yürütülmüştür. Puanlayıcılara ait D çalışması istatistikleri Tablo 11'de sunulmuştur.

Tablo 11

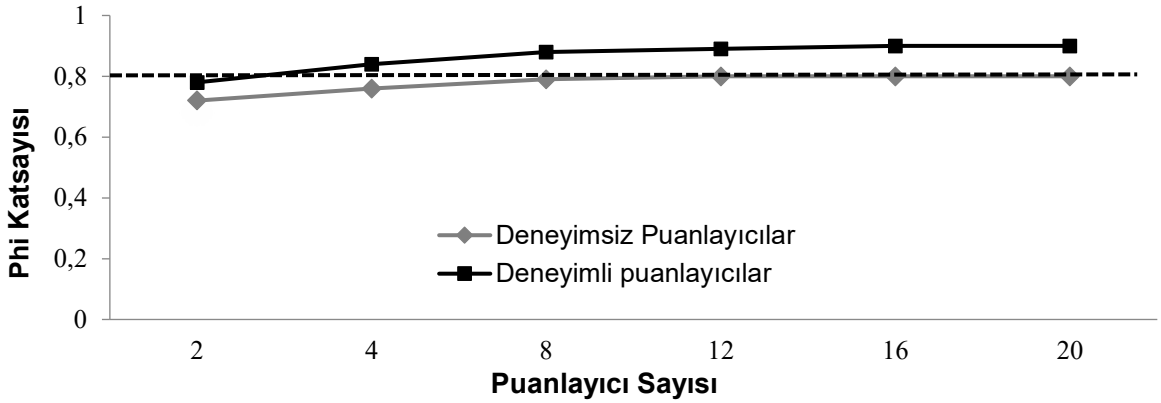
Sınıf Öğretmenleri İçin bxmxp Deseni D Çalışması Sonuçları

	Deneyimsiz						Deneyimli					
Puanlayıcı Sayısı	2	4	8	12	16	20	2	4	8	12	16	20
G Katsayısı	0,72	0,76	0,79	0,80	0,80	0,80	0,82	0,88	0,92	0,93	0,94	0,94
Phi Katsayısı	0,72	0,76	0,79	0,80	0,80	0,80	0,78	0,84	0,88	0,89	0,90	0,90
Bağıl Hata	0,073	0,058	0,050	0,047	0,046	0,045	0,072	0,042	0,027	0,022	0,019	0,017
Mutlak Hata	0,073	0,058	0,050	0,047	0,046	0,045	0,092	0,060	0,045	0,039	0,037	0,035



(Kesik çizgiler, yeterli genellenebilirlik ve güvenirlik için ≥ 0,80 kriterini göstermektedir.)

Şekil 5. D Çalışmalarına yönelik G katsayısı grafiği.



Şekil 6. D Çalışmalarına yönelik Phi katsayısı grafiği.

Puanlamalar için uygun değerlendirme prosedürleri kullanılarak yeterli genellenebilirlik ($E\rho^2$) ve güvenirliğin (Φ) (≥ 0,80) elde edilip edilemeyeceğini görmek araştırma açısından önemlidir. Bu nedenle, diğer yüzeyler mevcut seviyelerde sabit tutulurken puanlayıcı yüzeyinin sistematik olarak manipüle edildiği bir dizi D çalışması yapılmıştır. Bu sayede, puanlayıcı yüzeyinin genellenebilirlik ve güvenirliğe katkısının araştırılması amaçlanmıştır. Çalışmaların sonuçlarına göre puanlayıcı sayısının artmasının G ve Phi katsayılarını da artıracağı öngörülmektedir. Bu işlemlerin sonuçları Tablo 12, şekil 5 ve şekil 6'da gösterilmektedir. Puanlayıcı sayıları 4-8-12-16 ve son olarak 20 puanlayıcı olacak şekilde sistematik olarak değiştirilmiştir. Bu süreç, elde edilen G ve Phi katsayıları 0,80 kriterinin üzerinde bir değere çıkana ve nispeten sabit bir seriye ulaşana kadar sürdürülmüştür.

Birey ve madde sayılarının sabit tutulduğu ve puanlayıcı sayısının kademeli olarak artırıldığı senaryoda, deneysimsiz puanlayıcılar için 12 puanlayıcı (0,80) ile yeterli G ve Phi katsayıları elde edilmiştir. Puanlayıcı sayısının 16 ve 20 olduğu durumlarda katsayıların değişmediği gözlenmiştir. Deneyimli puanlayıcılar için G katsayısı 0,80 düzeyinin üzerinde olmakla birlikte Phi katsayısına ilişkin değerler incelendiğinde, 4 puanlayıcı için güvenilir sonuçlar elde edilebileceği (0,84) ve puanlamanın istenilen düzeyde yapılabileceği söylenebilir. Puanlayıcı sayısı 16 ve 20 olduğunda ise 0,90 gibi yüksek düzeyde Phi katsayısına ulaşılabilmektedir. Her iki puanlayıcı gurubunda puanlayıcı sayısının artırılması sonucu bağıl ve mutlak hata varyanslarının düştüğü gözlenmektedir.

Deneyimsiz puanlayıcılar için 2 puanlayıcının olduğu durumda bağıl ve mutlak hata varyansı 0,073 iken, puanlayıcı sayılarının 20 olduğu durumda bu değer 0,045'e düşmüştür. Deneyimli puanlayıcılarda 2 puanlayıcı için bağıl hata varyansı 0,072 ve mutlak hata varyansı 0,092 iken, puanlayıcı sayısının kademeli artırımı sonucu 20 puanlayıcı için 0,017 ve mutlak hata varyansı 0,035 düzeyine düşmüştür. Birey ve madde sayılarının sabit tutulduğu durumlarda puanlayıcı sayılarının artırılması G ve Phi katsayılarında istenilen düzey olan 0,80 seviyesini yakalamıştır. Araştırmanın amacı puanlayıcılar arası uyumun incelenmesi olduğu için madde ve birey sayılarında değişiklik yapılmamıştır. Tablo 12'de yeterli kabul edilecek güvenirlik sayılarına ulaşmak için gereken puanlayıcı sayıları yer almaktadır.

Tablo 12

Sınıf öğretmenlerinin puanlama deneyimlerine göre en uygun G ve Phi katsayılarının elde edildiği Puanlayıcı Sayıları

Puanlama Deneyimi	Puanlayıcı Sayısı	G katsayısı	Phi Katsayısı
Deneyimli	4	0,88	0,84
Deneyimsiz	12	0,80	0,80

Tablo 12 incelendiğinde en yüksek G ve Phi katsayılarına deneyimli puanlayıcılar için 4 puanlayıcı ve deneyimsiz puanlayıcılar için 12 puanlayıcı olduğu durumlarda ulaşıldığı görülebilir. Bununla birlikte deneyimsiz puanlayıcılar için kestirilen 12 sayısının uygulamada ulaşılabilir olmadığı yorumu yapılabilir. Hipotetik olarak yürütülen bu çalışma sonucunda deneyimli puanlayıcıların nispeten daha uyumlu puanlama yaptıkları düşünülebilir.

Alt Problem 2'ye Yönelik Bulgu ve Yorumlar

"Puanlama deneyimine sahip ve daha önce puanlama deneyimi olmayan Türkçe öğretmenlerinin puanlayıcılar arası uyum düzeyi nedir?" problemine yönelik Türkçe öğretmenlerinin puanlayıcılar arası uyum düzeyleri analizi bulguları; Cohen'in Ağırlıklandırılmış kappa, Krippendorff'un Alpha ve G kuramına göre, deneyim durumları göz önüne alınarak ayrı ayrı verilmiştir.

Tablo 13

7. sınıf Türkçe Öğretmenlerine Ait Puanlayıcılar Arası Uyum Düzeyleri

Görev									
		κ	κ (p)	α	α (q)	κ	κ (P)	α	α (q)
C1		0,607	0,000	0,726	0,207	0,712	0,000	0,784	0,040
C2		0,632	0,000	0,711	0,306	0,655	0,000	0,751	0,125
C3		0,177	0,035	0,175	1,0	0,605	0,000	0,632	0,664
C4		0,387	0,000	0,451	0,994	0,620	0,000	0,689	0,377
P1	Ö1	0,482	0,000	0,524	0,942	0,492	0,000	0,548	0,953
	Ö2	0,493	0,000	0,550	0,920	0,385	0,000	0,417	0,999
	Ö3	0,507	0,000	0,572	0,850	0471	0,000	0,585	0,882
	Ö4	0,367	0,000	0,467	0,986	0,500	0,000	0,579	0,811
P2	Ö1	0,598	0,000	0,206	1,0	0,401	0,000	0,301	1,0
	Ö2	0,678	0,000	0,801	0,003	0,380	0,000	0,398	0,999
	Ö3	0,696	0,000	0,816	0,003	0,415	0,000	0,488	0,995
	Ö4	0,492	0,000	0,654	0,566	0,517	0,000	0,664	0,504
P3	Ö1	0,594	0,000	0,760	0,078	0,142	0,001	0-,073	1,0
	Ö2	0,552	0,000	0,721	0,224	0,137	0,001	-0,037	1,0
	Ö3	0, 601	0,000	0,731	0,194	0,411	0,000	0,475	0,986
	Ö4	0,572	0,000	0,762	0,058	0,203	0,000	0,035	1,0
	Ö5	0,505	0,000	0,677	0,435	,259	0,000	0,153	1,0
M1	Ö1	0,259	0,001	0,279	0,999	0,411	0,000	0,487	0,967
	Ö2	0,335	0,000	0,302	0,999	0,539	0,000	0,599	0,751
	Ö3	0,279	0,001	0,339	0,999	0,290	0,000	0,272	1,0
	Ö4	0,131	0,075	0,187	1,0	0,148	0,000	-0,025	1,0
	Ö5	0,289	0,002	0,338	0,999	0,377	0,000	0,417	0,998
	Ö6	0,295	0,001	0,342	0,999	0,238	0,000	0,180	1,0
	Ö7	0,317	0,000	0,423	0,999	0,265	0,000	0,269	1,0
	Ö8	0,203	0,006	0,245	0,999	0,439	0,000	0,493	0,935

Tablo 13'te Yazma Becerileri Testini alan 7. sınıf öğrencilerinin verdiği yanıtlar doğrultusunda deneyimli ve deneyimsiz Türkçe öğretmenlerinden oluşan puanlayıcıların verdikleri puanların birbiriyle uyumunu incelemek amacıyla Cohen'in kappa (κ) ve Krippendorff'un Alpha (α) katsayılarına ait analiz sonuçları sunulmuştur.

Alt problem 2-a, Cohen'in ağırlıklandırılmış kappa istatistiğine göre puanlama deneyimine sahip ve daha önce puanlama deneyimi olmayan Türkçe öğretmenlerinin puanlayıcılar arası uyum düzeyi nedir? Deneyimsiz puanlayıcılara ait sütun incelendiğinde κ'ya göre en fazla uyum ikinci paragraf görevine ait üçüncü ölçüttür (P2Ö3, κ = 0,696). Söz konusu ölçütteki uyumun iyi derecede olduğu söylenebilir. Deneyimsiz puanlayıcılar arasında P2Ö3 ölçütünü sırasıyla P2Ö2 (0,678), C2 (0,632), C1(0,607) ve P3Ö3 (0,601) madde ve ölçütleri takip etmektedir. Söz konusu ölçüt ve maddelerdeki puanlayıcı uyum düzeyleri iyi olarak kabul edilmektedir. Deneyimsiz puanlayıcılar en yüksek uyumu ikinci paragraf görevi ile birinci ve ikinci cümle görevlerinde göstermişlerdir.

Deneyimsiz puanlayıcıların κ'ya göre mükemmel uyum gösterdiği bir madde veya ölçüt bulunmamakla birlikte en düşük uyum metin görevinin dördüncü ölçütünde gözlenmiştir (M1Ö4, κ=0,131). Bu ölçütü sırasıyla C3 (0,177), M1Ö8 (0,203), M1Ö1 (0,259) ve M1Ö3 (0,279) ölçüt ve maddeleri takip etmektedir. Dördüncü cümle görevi de κ = 0,387 ile zayıf uyumun gözlendiği bir diğer madde olmuştur. İlgili ölçüt ve maddelerdeki uyum düzeyi zayıftır ve puanlayıcıların şansa bağlı uyum göstermiş olma olasılığı yüksektir. Deneyimsiz puanlayıcılar en düşük uyumu metin görevine ait ölçütlerin tümünde 0,40 olan sınırın altında göstermiştir. Bunun sebebi ölçütlerin sınırlarının iyi belirlenememiş olması, yönergelerin yetersiz olması, puanlama anahtarının analitik olması veya kestirilemeyen değişkenlerin sürece karışmış olması gibi sebepler olabilir. Deneyimsiz puanlayıcılar için tüm görevlerde ortalama κ oranı 0,434 olarak bulunmuştur.

Tablo 14'de deneyimli puanlayıcılara ait sütun incelendiğinde κ'ya göre en yüksek uyum düzeyi κ = 0,712 ile birinci cümle görevinde (C1) gözlenmiştir. C1 görevindeki bu uyum düzeyi iyi olarak kabul edilmektedir. Deneyimli puanlayıcılar arasında C1 görevini sırasıyla C2 (0,655), C4 (0,620), C3 (0,605) maddeleri ile M1Ö1 (0,539), P2Ö4 (0,517) ve P1Ö4 (0,500) ölçütleri takip etmektedir. Cümle

görevlerinin tümünde iyi derecede uyum gözlenen deneyimli puanlayıcılar, M1Ö1, P2Ö4 ve P1Ö4 ölçütlerinde de yeterli uyum düzeyine ulaşmışlardır.

Deneyimli puanlayıcılar arasında κ'ya göre en düşük uyumun gözlendiği ölçüt üçüncü paragraf görevinin ikinci ölçütüdür. (P3Ö2, κ = 0,137). Bu ölçütü sırasıyla P3Ö1 (0,142), M1Ö4 (0,148), P3Ö4 (0,203), M1Ö6 (0,238) ve P3Ö5 (0,259) ölçütleri takip etmektedir ve bu uyum düzeyleri çok zayıf olarak nitelendirilebilir. Üçüncü paragraf görevine dikkat edilirse, deneyimli puanlayıcılar bu görevin neredeyse tüm ölçütlerinde çok zayıf uyum göstermiştir. Aynı görevde deneyimsiz puanlayıcılar nispeten daha yüksek uyum düzeyine sahiptir. Deneyimsiz puanlayıcılar ile kıyaslandığında cümle görevlerinde daha yüksek uyum düzeyine sahip olan deneyimli puanlayıcıların tüm görevlerdeki κ uyum düzeyi ortalaması 0,400 olarak hesaplanmıştır ve bu ortalama deneyimsiz puanlayıcıların uyum ortalamasına göre (0,434) daha düşüktür.

Alt problem 2-b, Krippendorff'un Alpha katsayısı istatistiğine göre puanlama deneyimine sahip ve daha önce puanlama deneyimi olmayan Türkçe öğretmenlerinin puanlayıcılar arası uyum düzeyi nedir? Tablo 14'te bulunan maddelere ait Krippendorff α katsayıları incelendiğinde deneyimsiz puanlayıcılar arasında en yüksek uyumun gözlendiği ölçüt 0,816 ile ikinci paragraf görevinin üçüncü ölçütüdür (P2Ö3). Bu ölçütü sırasıyla P2Ö2 (0,801), P3Ö4 (0,762), P3Ö1 (0,760), P3Ö3 (0,731), P3Ö2 (0,721), P3Ö5 (0,677) ölçütleri ile C1 (0,726) ve C2 (0,711) görevleri takip etmektedir. 0,67 değerinin altındaki uyumun zayıf olarak değerlendirildiği göz önüne alındığında, P2Ö3 ve P2Ö2 ölçütlerinde yüksek uyum, P2Ö2, P3Ö4, P3Ö1, P3Ö3, P3Ö2 ve P3Ö5 ölçütlerinde orta derecede uyum ve aynı zamanda C1 ve C1 görevlerinde de orta derecede uyum gözlenmiştir. Bunların dışında kalan ölçüt ve maddelerdeki puanlayıcı uyumları genel olarak zayıf kalmıştır.

En zayıf uyumun olduğu madde κ'da olduğu gibi α=0,175 değeriyle C3 maddesidir. Bu maddeyi sırasıyla M1Ö4 (0,187), P2Ö1 (0,361), M1Ö8 (0,245), M1Ö2 (0,302) ve M1Ö3 (0,339) ölçütleri takip etmiştir. Genel olarak bakıldığında en düşük uyum düzeyinin metin görevine ait ölçütlerde olduğu söylenebilir. Olasılık (q) indeksi incelendiğinde, C3 maddesi ile P2Ö1, M1Ö4, M1Ö5, M1Ö6, M1Ö7 ve M1Ö8 ölçütleri, tüm popülasyonun test edilmesi durumunda neredeyse %100 olasılıkla en düşük uyum değeri olan 0,67 düzeyinin altında kalacaktır. Bu düzeyin üzerinde olma

ihtimali en yüksek olan ölçütler %3 ile paragraf görevinin ikinci ve üçüncü ölçütüdür. Deneyimsiz 7. sınıf Türkçe öğretmenlerinin gerçekleştirdiği puanlama süreci sonucunda tüm madde ve ölçütlerdeki ortalama Krippendorf Alpha değeri 0,510 olarak hesaplanmıştır

Tablo 14'te deneyimli puanlayıcılara ait Krippendorff α katsayıları incelendiğinde, en yüksek uyum düzeyi 0,784 ile birinci cümle görevidir. Bu görevi sırasıyla C2 (0,751), C4 (0,689), P2Ö4 (0,664) ve C3 (0,632) görev ve ölçütleri takip etmektedir. C1, C2 ve C4 görevler orta düzeyde uyuma sahipken puanlayıcılar diğer ölçüt ve görevlerin tümünde zayıf uyum ile puanlama yapmışlardır.

En zayıf uyumun olduğu ve negatif değere sahip P3Ö1 (-,073), P3Ö2 (-,037) ve M1Ö4 (-0,25) ölçütleri, şans uyumunun altında bir uzlaşmaya işaret etmekte ve bu üç ölçüt için sürekli bir uyumsuzluğun ortaya çıktığı söylenebilir. En düşük uyum düzeyleri üçüncü paragraf görevi ölçütleri olan P3Ö1 ve P3Ö2 ile metin görevine ait dördüncü ölçüt olan M1Ö4'tür.ir. Olasılık (q) indeksleri incelendiğinde, 3. Paragraf görevine ait tüm ölçütlerde, tüm popülasyonda yapılacak bir test uygulamasında uyum düzeyinin 0,67 olan sınır değerinin altında olma ihtimali neredeyse %100 olarak bulunmuştur. Deneyimli 7. sınıf Türkçe öğretmenlerinin gerçekleştirdiği puanlama süreci sonucunda tüm madde ve ölçütlerdeki ortalama Krippendorff Alpha değeri 0,40324 olarak hesaplanmıştır

7. sınıf Türkçe öğretmenlerine yönelik yapılan uyum düzeyleri karşılaştırması sonucunda genel olarak deneyimsiz puanlayıcılar deneyimli puanlayıcılara göre, tüm görevlerdeki κ ve α düzeyleri ortalamasına göre daha uyumlu puanlama gerçekleştirmiştir. Deneyimsiz öğretmenler en yüksek uyum düzeyini ikinci ve üçüncü paragraf görevlerine ait ölçütlerde (P2 ve P3) gösterirken deneyimli öğretmenler en yüksek uyum düzeyini cümle görevlerinde göstermiştir. Dolayısıyla deneyimli öğretmenlerin bütünsel DPA kullanılan görevlerde daha yüksek uyum düzeyine sahip olduğu söylenebilir. Bununla birlikte deneyimsiz Türkçe öğretmenleri analitik DPA'lar ile puanlanan görevlerde nispeten daha yüksek uyum göstermiştir.

Deneyimsiz Türkçe öğretmenlerinin uyum düzeylerinin en düşük olduğu görev üçüncü cümle (C3) görevi iken, ölçüt bazında en düşük uyum düzeyleri metin görevine ait ölçütlerdedir. Deneyimli Türkçe öğretmenlerine gelindiğinde ise, en düşük uyum düzeylerinin üçüncü paragraf görevine (P3) ait ölçütlerde görüldüğü

söylenebilir. Bunun yanında, metin görevine ait ölçütlerde de düşük uyumun gözlendiği görülmektedir. Her iki grup, metin görevine ait ölçütlerde düşük uyum göstermiştir.

Alt problem 2-c, G kuramına göre puanlama deneyimine sahip ve daha önce puanlama deneyimi olmayan Türkçe öğretmenlerinin G ve phi katsayısı düzeyleri nedir? İkinci alt probleme ait son soru olan G kuramına yönelik analizler gerçekleştirilmiştir. Bu analiz sonuçlarının sunulmasından önce oluşturulan desen ve yüzeylere ilişkin bilgiler Tablo 14'de sunulmuştur.

Tablo 14

İkinci Alt Probleme Ait G Kuramı Deseninde Bulunan Yüzeyler

Yüzey	Etiket	Düzey	Evren
Birey	B	80	SONSUZ
Madde	M	25	SONSUZ
Puanlayıcı	P	2	SONSUZ

G kuramı altında yapılan analizlere yönelik tümüyle çaprazlanmış bxmxp desenine ilişkin varyans bileşenleri bulguları Tablo 15'te gösterilmiştir.

Tablo 15

Türkçe Öğretmenlerinin Tümüyle Çaprazlanmış bxmxp Desenine Ait Kestirilen Varyans Bileşenleri

Varyans Kaynağı	Varyans	% Varyans	Varyans	% Varyans
B	0,41298	25,0	0,35030	22,8
M	0,09121	5,5	0,00947	0,6
P	0,00231	0,1	0,12422	8,1
BM	0,46721	28,3	0,45959	29,9
BP	0,05917	3,6	0,06373	4,2
MP	0,02799	1,7	0,07559	4,9

BMP	0,59278	35,8	0,45170	29,4
Toplam		%100		%100
$E\rho^2$		0,87		0,86
Φ		0,86		0,74

Tablo 15'deki deneyimsiz puanlayıcılara ait sütun incelendiğinde, toplam varyansın %0,1'i puanlayıcılardan kaynaklanan etkilere atfedilmiştir. Bu, toplam değişkenliğin neredeyse hiçbirinin puanlayıcılara atfedilmediği anlamına gelebilir. Birey ve puanlayıcı (BP) (%3,6) ve madde ve puanlayıcı (%1,7) ortak etkilerine yönelik varyans kaynakları, puanlayıcı (%0,1) kaynaklı varyans ile birlikte toplam varyansa en az katkı sağlayan etkiler olarak görülebilir. Bu sonuca dayanarak, deneyimsiz Türkçe öğretmenlerinin bireyleri ve görevleri sıralamada büyük farklılıklar göstermedikleri söylenebilir. Bunun yanında, bireyler arası farklılıklardan kaynaklanan varyans yüzdesinin birey, madde ve puanlayıcı (BMP) ortak etkisine (%35,8) bağlı olduğu görülmektedir. Bu, ölçülemeyen değişkenlik ve hata kaynaklarının fazla olduğuna işaret etmektedir. Toplam varyans içindeki en düşük paya sahip olan puanlayıcı (P) yüzeyine atfedilebilen varyans yüzdesi (%0), öğretmenlerin gerçekleştirdiği puanlamaların genel olarak önemli ölçüde değişmediğini, katılık veya hoşgörü göstermediğine işaret etmektedir. Birey ve puanlayıcı (BP) arasındaki etkileşimin, gözlemlenen varyansın %3,6'sını açıkladığı göz önüne alındığında, öğretmenlerin belirli öğrencilere yönelik puanlamalarının bir miktar değişmekle birlikte bu değişimin nispeten düşük olduğu yorumu yapılabilir. Madde ve puanlayıcı (MP) arasındaki etkileşim, toplam varyansa %1,7 civarında katkıda bulunmuştur. Bu varyans yüzdesinin küçük olması, maddeler arasındaki puanlama davranışında çok küçük değişiklikler olduğunu göstermektedir. Ölçme objesi olan bireylerin toplam varyans içerisindeki oranı üçüncü en büyük varyans oranı olan %25'tir. Bireylerden kaynaklanan varyans oranının büyük olması genelde istenen bir durumdur. Bu puanlayıcıların öğrencilerin arasındaki farklılıkları ortalama düzeyde ayırabildiğini gösterir. Son olarak, madde yüzeyinin (M) toplam varyans içinde çok büyük bir payı olmadığı söylenebilir. (%23,4). Bu, maddelerin bir maddeden diğerine zorluk ve ayırt edicilik olarak değişkenlik göstermediği şeklinde yorumlanabilir.

Tablo 15'deki deneyimli puanlayıcılara ait sütun incelendiğinde, toplam varyansın %8,1'i puanlayıcılardan (P) kaynaklanan etkilerle ilişkilidir. Deneyimsiz puanlayıcılarla bu oranın neredeyse önemsiz (%0,1) derecede olduğu düşünüldüğünde, deneyimli puanlayıcıların nispeten daha katı/cömert puanlama yaptıkları söylenebilir. Bireylerden kaynaklanan varyans (B), neredeyse %23 civarındadır ve toplam varyans içinde en büyük üçüncü paya sahiptir. Bireyler için kestirilen varyans değeri payının büyüklüğü, bireyler arasındaki sistematik farklılıkların ortaya konulduğunu göstermektedir. Bu bağlamda her ne kadar en büyük pay bu varyans bileşenine ait olmasa da belirli bir oranda birey farklılıklarının ayrıldığı söylenebilir. Görevlerden (M) kaynaklanan değişkenlik etkisinin dikkate alınmayacak kadar az olduğu görülmektedir (%0,6). Bu oran doğrultusunda görev güçlük düzeylerinin benzeşik olduğu söylenebilir. Birey-madde ortak etkisi (BM) incelendiğinde, toplam varyans içindeki en büyük payın neredeyse %30 ile bu etkiden kaynaklandığı söylenebilir. Bu durum, belirli bireylerin bir görevden diğerine bağıl durumlarındaki farklılıklarının arttığı şeklinde yorumlanabilir. Birey-puanlayıcı (BP, %4,2) ve madde-puanlayıcı (MP, %4,9) ortak etkileri incelendiğinde atfedilen varyans oranlarının nispeten düşük olduğu görülmektedir. Bu durum, belirli puanlayıcıların belirli bireyler arasındaki puanlamalarının çok az değişiklik gösterdiği ve puanlayıcıların görevler arasında nispeten kararlı puanlamalar yaptıkları şeklinde yorumlanabilir. Son olarak, en büyük ikinci varyans yüzdesinin artık hata ile birlikte anılan birey, madde ve puanlayıcı (BMP) ortak etkisine (%29,4) bağlı olduğu görülmektedir. Bu, ölçülemeyen değişkenlik kaynaklarının çokluğuna işaret etmektedir.

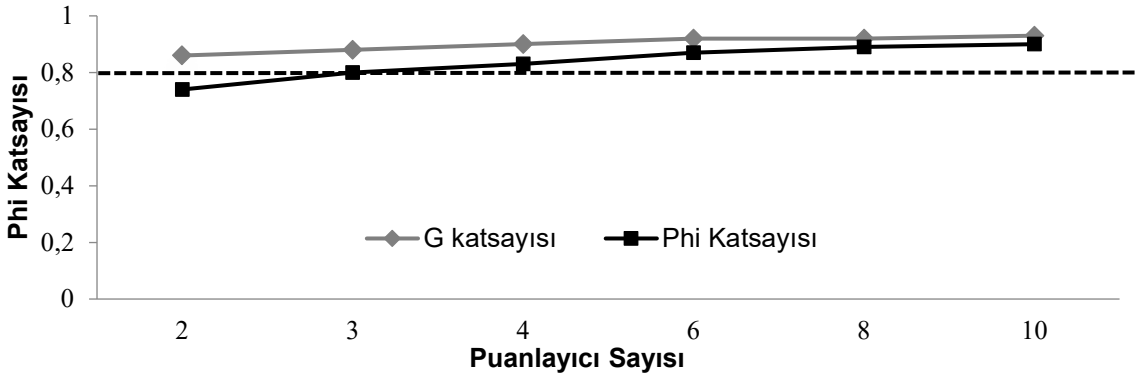
Deneyimsiz puanlayıcılara yönelik yapılan G çalışması sonucu ölçme objesine bağlı varyans bileşenlerinin toplamı sonucu elde edilen bağıl hata varyansı 0,06013, G katsayısı 0,87 ve Phi katsayısı 0.86 olarak kestirilmiştir. Kestirilen G ve Phi katsayılarının istenen düzeyde olduğu görülmektedir. Deneyimsiz 7. Sınıf Türkçe öğretmenleri, Yazma Becerileri Testinde, G kuramına göre, aşırı katı veya hoşgörülü olmadan ve bireyler ile görevler arasında değişiklik göstermeden puanlama gerçekleştirmişlerdir. Bununla birlikte kaynağı belli olmayan etkilerin çok olduğu görülmektedir. Deneyimli puanlayıcılara yönelik gerçekleştirilen G çalışması sonucu, bağıl hata varyansını 0,05928, G katsasyısı $E\rho^2$ =0,86 ve Phi katsayısı da Φ =0,74 olarak bulunmuştur. Kestirilen G değeri istenilen seviyede olmakla birlikte

Phi değerinin istenilen (≥ .80) değerin altında olduğu görülmektedir. İstenilen Phi düzeyine ulaşmak için puanlayıcı sayılarının değişimlendiği bir D çalışması yürütülmüştür. İlgili D çalışması istatistikleri tablo 16'da sunulmuştur. Ayrıca, G kuramı altında yapılan analiz sonuçlarına göre, puanlama deneyimi olmayan Türkçe öğretmenlerinin yaptığı puanlamalara yönelik G katsayısı (0,87) ve Phi katsayısı (0,86), puanlama deneyimi olan Türkçe öğretmenlerine ait G (0,86) ve Phi (0,74) katsayılarına göre daha yüksek bulunmuştur.

Tablo 16

Puanlama Deneyimi Olan Türkçe Öğretmenleri İçin bxmxp Deseni D Çalışması Sonuçları

Puanlayıcı Sayısı	2	3	4	6	8	10
G Katsayısı	0,86	0,88	0,90	0,92	0,92	0,93
Phi Katsayısı	0,74	0,80	0,83	0,87	0,89	0,90
Bağıl Hata	0,059	0,045	0,038	0,032	0,028	0,026
Mutlak Hata	0,123	0,088	0,071	0,053	0,044	0,039



Şekil 7. Puanlama deneyimine sahip türkçe öğretmenleri için d çalışmalarına yönelik G ve Phi katsayıları grafiği.

Yeterli G katsayısı düzeyine ulaşılmakla birlikte, puanlamalar için uygun değerlendirme prosedürleri kullanılarak yeterli güvenirliğin(Φ) (≥ 0,80) elde edilip

edilemeyeceğini görmek amacıyla diğer yüzeylerde bulunan görev ve birey sayıları sabit tutulurken puanlayıcı yüzeyinin sistematik olarak değiştirildiği bir D çalışması yapılmıştır. Bu sayede, puanlayıcı yüzeyinin genellenebilirlik ve güvenirliğe katkısının araştırılması amaçlanmıştır. Çalışmaların sonuçlarına göre puanlayıcı sayısının artmasının G ve Phi katsayılarını da artırdığı görülmüştür. Bu işlemlerin sonuçları Tablo 16 ve Şekil 7'de gösterilmektedir. Puanlayıcı sayıları 3-4-6-8 ve son olarak 10 puanlayıcı olacak şekilde sistematik olarak değiştirilmiştir. Bu süreç, elde edilen Phi katsayısı 0,80 kriterinin üzerinde bir değere çıkana ve sabit bir seriye ulaşana kadar sürdürülmüştür.

Birey ve madde sayılarının sabit tutulduğu ve puanlayıcı sayısının kademeli olarak artırıldığı senaryoda, deneysimli puanlayıcılar için 3 puanlayıcı (0,80) ile yeterli Phi katsayısı elde edilmiştir. Puanlayıcı sayısının 8 ve 10 olduğu durumlarda katsayıların çok az bir değişiklik gösterdiği görülebilir.

Deneyimsiz puanlayıcılar için 2 puanlayıcının olduğu durumda bağıl ve mutlak hata varyansı sırasıyla 0,059 ve 0,123 iken, puanlayıcı sayılarının 10 olduğu durumda bu değer sırasıyla 0,026 ve 0,039'a düşmüştür. Birey ve madde sayılarının sabit tutulduğu durumlarda puanlayıcı sayılarının artırılması sonucunda Phi katsayısı bağlamında istenilen düzey olan 0,80 seviyesini yakalamıştır. Tablo 17'de en uygun güvenirlik sayılarına ulaşmak için gereken puanlayıcı sayıları yer almaktadır.

Tablo 17

Deneyimli Türkçe Öğretmenleri İçin En Uygun G ve Phi Katsayılarının Elde Edildiği Puanlayıcı Sayıları

Puanlayıcı Sayısı	G katsayısı	Phi Katsayısı
3	0,88	0,80

Tablo 17 incelendiğinde deneyimli puanlayıcılar için en yüksek G ve Phi katsayılarına puanlayıcı sayısının 3'e çıktığı durumlarda ulaşıldığı görülebilir. Puanlayıcı sayısının artması her kadar güvenirliği artırsa da, belirli bir puanlayıcı sayısının üzerine çıkılması uygulamada pratik olmaktan uzaklaşabileceği düşünülebilir.

Alt Problem 3'e Yönelik Bulgu ve Yorumlar

"Puanlama deneyimine sahip ve daha önce puanlama deneyimi olmayan Türk dili ve edebiyatı öğretmenlerinin puanlayıcılar arası uyum düzeyi nedir?" problemine yönelik Türk dili ve edebiyatı öğretmenlerinin puanlayıcılar arası uyum düzeyleri analizi bulguları; Cohen'in Ağırlıklandırılmış kappa, Krippendorff'un Alpha ve G kuramına göre, deneyim durumları göz önüne alınarak ayrı ayrı verilmiştir.

Tablo 18

9. sınıf Türk Dili ve Edebiyatı Öğretmenlerine Ait Puanlayıcılar Arası Uyum Düzeyleri

Görev		κ	κ (p)	α	α (q)	κ	κ (P)	α	α (q)
C1		0,694	0,000	0,733	0,170	0,639	0,000	0,719	0,219
C2		0,408	0,000	0,473	0,984	0,264	0,001	0,369	1,0
C3		0,558	0,000	0,623	0,747	0,595	0,000	0,683	0,383
C4		0,584	0,000	0,641	0,645	0,508	0,000	0,575	0,869
P1	Ö1	0,323	0,000	0,337	0,999	0,654	0,000	0,780	0,024
	Ö2	0,225	0,006	0,283	0,999	0,677	0,000	0,788	0,015
	Ö3	0,289	0,001	0,318	0,999	0,706	0,000	0,775	0,050
	Ö4	0,332	0,000	0,333	0,998	0,534	0,000	0,693	0,355
P2	Ö1	0,413	0,000	0,257	1,0	0,588	0,000	0,587	0,819
	Ö2	0,340	0,000	0,455	0,985	0,601	0,000	0,752	0,106
	Ö3	0,373	0,000	0,472	0,977	0,622	0,000	0,748	0,140
	Ö4	0,449	0,000	0,503	0,952	0,559	0,000	0,714	0,255
P3	Ö1	0,437	0,000	0,539	0,937	0,436	0,000	0,507	0,946
	Ö2	0,461	0,000	0,507	0,955	0,382	0,000	0,517	0,940
	Ö3	0,408	0,000	0,497	0,974	0,461	0,000	0,537	0,921
	Ö4	0,609	0,000	0,669	0,481	0,389	0,000	0,490	0,959
	Ö5	0,513	0,000	0,610	0,775	0,356	0,000	0,462	0,972

M1	Ö1	0,712	0,000	0,813	0,003	0,553	0,000	0,726	0,150
	Ö2	0,586	0,000	0,668	0,474	0,395	0,000	0,490	0,994
	Ö3	0,665	0,000	0,772	0,023	0,528	0,000	0,674	0,447
	Ö4	0,677	0,000	0,755	0,003	0,298	0,000	0,360	1,0
	Ö5	0,581	0,000	0,692	0,323	0,293	0,000	0,362	1,0
	Ö6	0,352	0,000	0,397	1,0	0,258	0,000	0,342	1,0

Tablo 18'de Yazma Becerileri Testini alan 9. sınıf öğrencilerinin verdiği yanıtlar doğrultusunda deneyimli ve deneyimsiz Türk dili ve edebiyatı öğretmenlerinden oluşan puanlayıcıların verdikleri puanlar arasındaki uyumun incelenmesi amacıyla yapılan Cohen'in kappa (κ) ve Krippendorff'un Alpha (α) katsayılarına ait analiz sonuçları sunulmuştur.

Alt problem 3-a, Cohen'in ağırlıklandırılmış kappa istatistiğine göre puanlama deneyimine sahip ve daha önce puanlama deneyimi olmayan Türk dili ve edebiyatı öğretmenlerinin puanlayıcılar arası uyum düzeyi nedir? Deneyimsiz puanlayıcılara ait sütun incelendiğinde κ'ya göre en fazla uyum metin görevinin birinci ölçütünde gözlenmiştir. (M1Ö1, κ = 0,712). Ölçütteki puanlayıcılar arası uyum iyi düzeyde kabul edilmektedir. Deneyimsiz puanlayıcılar arasında M1Ö1 ölçütünü sırasıyla birinci cümle görevi C1 (0,694), metin görevinin dördüncü, üçüncü, ikinci ve beşinci ölçütleri M1Ö4 (0,677), M1Ö3 (0,665), M1Ö2 (0,586), M1Ö5 (0,581) ve dördüncü cümle görevi C4 (0,584) takip etmektedir. M1Ö4 ve M1Ö3 ölçütlerindeki uyum düzeyi iyi ve M1Ö2, M1Ö5 ölçütleri ile C4 görevindeki uyum düzeyleri de yeterli düzeyde kabul edilmektedir. Deneyimsiz puanlayıcılar en yüksek uyumu metin görevine ait ölçütlerde göstermişlerdir. Diğer sınıf düzeyleri ile kıyaslandığında metin görevindeki bu nispeten yüksek uyumun yüksek olduğu söylenebilir.

Deneyimsiz puanlayıcıların κ istatistiğine göre üst düzey uyumluluk gösterdiği bir madde veya ölçüt bulunmamakla birlikte en düşük uyum birinci paragraf görevinin ikinci ölçütünde gözlenmiştir (P1Ö2, κ = 0,225). Bu ölçütü sırasıyla P1Ö3 (0,289), P1Ö1 (0,323), P1Ö4 (0,332) ve P1Ö2 (0,340) ölçütleri takip etmektedir. Bu ölçütlerdeki uyum düzeyinin zayıf olduğu ve puanlayıcıların şansa bağlı uyum göstermiş olma ihtimallerinin yüksek olduğu söylenebilir. Deneyimsiz puanlayıcılara

ait κ istatistiği incelendiğinde en düşük uyumun birinci paragraf görevine ait ölçütlerde gözlendiği görülmektedir. Bu ölçütlerin tümünde 0,40 olan sınırın altında bir uyum düzeyi söz konusudur. Deneyimsiz puanlayıcılar için tüm görev ve ölçütlerde ortalama κ oranı 0,477 olarak bulunmuştur.

Tablo 18'de deneyimli puanlayıcılara ait sütun incelenirse κ'ya göre en yüksek uyum düzeyinin κ = 0,706 ile birinci paragraf görevinin üçüncü ölçütünde (P1Ö3) görüldüğü söylenebilir. P1Ö3 ölçütündeki bu uyum düzeyi iyi olarak kabul edilmektedir. Deneyimli puanlayıcılar arasında P1Ö3 ölçütünü sırasıyla P1Ö2 (0,677), P1Ö1 (0,654), C1 (0,639), P2Ö3 (0,622) ve P2Ö2 (0,601) madde ve ölçütleri takip etmektedir. Deneyimsiz puanlayıcıların aksine, deneyimli Türk dili ve edebiyatı öğretmenleri birinci paragraf görevinde nispeten yüksek uyum göstermiştir. Bunun yanında, deneyimli Türk dili ve edebiyatı öğretmenleri; C3, P2Ö1, P2Ö4, M1Ö1, M1Ö3, P1Ö4, P3Ö3 ve P3Ö1 ölçüt ve maddelerinde de yeterli uyum düzeyine ulaşmışlardır.

Deneyimli puanlayıcılar arasında κ'ya göre en düşük uyumun gözlendiği ölçüt üçüncü metin görevinin altıncı ölçütüdür (M1Ö6, κ = 0,258). Bu ölçütü sırasıyla C2 (0,264), M1Ö5 (0,293), M1Ö4 (0,298), P3Ö2 (0,382) ve P3Ö5 (0,356) ölçüt ve maddeleri takip etmektedir ve bu uyum düzeyleri zayıf olarak nitelendirilmektedir. Maddeler içinde en çok uyumsuzluğun metin görevinde olduğu söylenebilir. Bu görevdeki altı ölçütün dördü puanlayıcılar arasında zayıf bir uyum düzeyine sahiptir. Aynı görevin deneyimsiz puanlayıcılara ait istatistikleri incelendiğinde daha yüksek bir uyum düzeyinden söz edilebilir. Deneyimsiz puanlayıcılar ile kıyaslandığında birinci ve ikinci paragraf (P1,P2) görevlerinde daha yüksek uyum düzeyine sahip olan deneyimli puanlayıcıların tüm görevlerdeki κ uyum düzeyi ortalaması 0,491 olarak hesaplanmıştır ve bu ortalama deneyimsiz puanlayıcıların uyum ortalamasına göre (0,477) daha yüksek olmakla birlikte manidar bir büyüklükten söz etmek güçtür.

Alt problem 3-b, Krippendorff'un Alpha katsayısı istatistiğine göre puanlama deneyimine sahip ve daha önce puanlama deneyimi olmayan Türk dili ve edebiyatı öğretmenlerinin puanlayıcılar arası uyum düzeyi nedir? Tablo 18'de bulunan maddelere ait Krippendorff α katsayıları incelendiğinde deneyimsiz puanlayıcılar arasında en yüksek uyumun gözlendiği ölçüt 0,813 ile metin görevinin birinci ölçütüdür (M1Ö1). Bu ölçütü sırasıyla M1Ö3 (0,772), M1Ö4 (0,755), C1

(0,733), M1Ö5 (0,692), P3Ö4 (0,669) ve M1Ö2 (0,668) ölçüt ve maddeleri takip etmektedir. Dikkat edilecek olursa metin görevindeki altı ölçütten beşinin en az yeterli düzeyde olduğu söylenebilir. Buradan hareketle, deneyimsiz Türk dili ve edebiyatı öğretmenlerinin metin görevine ait ölçütlerde nispeten iyi uyum yakaladığı sonucuna ulaşılabilir. Bunların dışında kalan ölçüt ve maddelerdeki puanlayıcı uyumları zayıf olarak nitelendirilebilir.

En zayıf uyum α =0,257 değeriyle P2Ö1 ölçütünde gözlenmiştir. Bu ölçütü sırasıyla P1Ö2 (0,283), P1Ö3 (0,318), P1Ö4 (0,333), P1Ö1 (0,337) ve M1Ö6 (0,397) ölçütleri takip etmiştir. Genel olarak bakılacak olursa birinci paragraf görevinin tüm ölçütlerinde en düşük uyum değerleri gözlenmiştir. Bunun sebebi yönergenin net anlaşılmamış olabileceği veya madde ile ilgili bir problemden kaynaklı olabilir. Olasılık (q) indeksi incelendiğinde, C2 maddesi, birinci paragraf görevi ile ikinci paragraf görevinin tüm ölçütleri ve P3Ö1, P3Ö2 ve P3Ö3 ölçütlerin, tüm popülasyonun test edilmesi durumunda neredeyse %100 olasılıkla en düşük uyum değeri olan 0,67 düzeyinin altında kalacaktır. Deneyimsiz 9. sınıf Türk dili ve edebiyatı öğretmenlerinin gerçekleştirdiği puanlama süreci sonucunda tüm madde ve ölçütlerdeki ortalama Krippendorff Alpha değeri 0,536 olarak hesaplanmıştır

Tablo 18'de deneyimli puanlayıcılara ait Krippendorff α katsayıları incelendiğinde, en yüksek uyum düzeyi 0,788 ile birinci paragraf görevinin ikici ölçütünde gözlenmiştir. Bu ölçütü sırasıyla P1Ö1 (0,78), P1Ö3 (0,775), P2Ö2 (0,752), P2Ö3 (0,748), M1Ö1 (0,726), C1 (0,719), P2Ö4 (0,714), P1Ö4 (0,693), C3 (0,683), ve M1Ö3 (0,674) görev ve ölçütleri takip etmektedir. Bunların dışında kalan görev ve ölçütler minimum uyum düzeyi olan 0,67 değerinin altında kalarak yetersiz olarak nitelendirilebilir.

En zayıf uyum M1Ö6 (0,342) ölçütünde gözlenmekle birlikte bu ölçütü M1Ö4 (0,36), M1Ö5 (0,362) ölçütleri ve C2 (0,369) görevi takip etmektedir. Olasılık (q) indeksleri incelendiğinde, 3. Paragraf görevine ait tüm ölçütlerde, tüm popülasyonda yapılacak bir test uygulamasında uyum düzeyinin 0,67 olan sınır değerinin altında olma ihtimali neredeyse %100 olarak bulunmuştur. Deneyimli 9. sınıf Türk dili ve edebiyatı öğretmenlerinin gerçekleştirdiği puanlama süreci sonucunda tüm madde ve ölçütlerdeki ortalama Krippendorff Alpha değeri 0,593 olarak hesaplanmıştır. Deneyimli puanlayıcılar genel olarak birinci paragraf görevinde yüksek uyum göstermişken deneyimsiz puanlayıcılara tam aksine

uyumun en düşük olduğu görev bu olmuştur. Bu durumun bir sebebi yönergenin iki grup tarafından farklı anlaşılmış olabileceğidir.

9. sınıf Türk dili ve edebiyatı öğretmenlerine yönelik yapılan KTK'ya dayalı uyum düzeyleri karşılaştırması sonucunda cümle görevleri, üçüncü paragraf görevi ve metin görevlerinde deneyimsiz puanlayıcılardaki uyum düzeyi nispeten daha yüksek bulunmuşken, birinci ve ikinci paragraf görevlerinde deneyimli puanlayıcılara ait uyum düzeyleri daha üst seviyededir. Bu sonuçlara dayanarak DPA'ların türüne göre kullanımına ilişkin belirgin bir fark olduğunu söylemek güçtür.

Deneyimsiz Türk dili ve edebiyatı öğretmenlerinin uyum düzeylerinin en düşük olduğu görev birinci paragraf (P1) görevidir. Deneyimli Türk dili ve edebiyatı öğretmenlerine ait uyum düzeyleri incelendiğinde ise, en düşük uyum düzeylerinin üçüncü metin görevine (M) ait ölçütlerde görüldüğü söylenebilir.

Alt problem 3-c, G kuramına göre puanlama deneyimine sahip ve daha önce puanlama deneyimi olmayan Türk dili ve edebiyatı öğretmenlerinin G ve phi katsayısı düzeyleri nedir? Üçüncü alt probleme ait son soru olan G kuramına yönelik analizler gerçekleştirilmiştir. Bu analiz sonuçlarının sunulmasından önce oluşturulan desen ve yüzeylere ilişkin bilgiler Tablo 19'da sunulmuştur.

Tablo 19

Üçüncü Alt Probleme Ait G Kuramı Deseninde Bulunan Yüzeyler

Yüzey	Etiket	Düzey	Evren
Birey	B	80	SONSUZ
Madde	M	23	SONSUZ
Puanlayıcı	P	2	SONSUZ

G kuramı altında yapılan analizlere yönelik tümüyle çaprazlanmış bxmxp desenine ilişkin varyans bileşenleri bulguları Tablo 20'de gösterilmiştir.

Tablo 20

Türk Dili ve Edebiyatı Öğretmenlerinin Tümüyle Çaprazlanmış bxmxp Desenine Ait Kestirilen Varyans Bileşenleri

Varyans Kaynağı	Varyans	% Varyans	Varyans	% Varyans
B	0,45060	20,4	0,34334	17,5
M	0,20068	9,1	0,21787	11,1
P	-0,00330	0,0	-0,00127	0,0
BM	0,67380	30,6	0,74932	38,2
BP	0,09705	4,4	0,06641	3,4
MP	0,07531	3,4	0,00940	0,5
BMP	0,70676	32,1	0,57491	29,3
Toplam		100%		%100
$E\rho^2$	0,83		0,81	
Φ	0,81		0,80	

Tablo 20'deki deneyimsiz puanlayıcılara ait sütun incelendiğinde, puanlayıcılardan kaynaklanan etkilere atfedilen varyansın toplam varyans içindeki payının %0 olduğu görülmektedir. Bu, toplam değişkenliğin neredeyse hiçbirinin puanlayıcılara atfedilmediği şeklinde yorumlanabilir. Birey ve puanlayıcı (BP) (%4,4) ve madde ve puanlayıcı (%3,4) ortak etkilerine yönelik varyans kaynakları, puanlayıcı (%0) kaynaklı varyans ile birlikte toplam varyansa en az katkı sağlayan etkilerdir. Bu verilere dayanarak, deneyimsiz Türk dili ve edebiyatı öğretmenlerinin bireyleri ve görevleri sıralamada birbirinden büyük farklılıklar göstermedikleri söylenebilir. Bunun yanında, bireyler arası farklılıklardan kaynaklanan varyans yüzdesinin birey, madde ve puanlayıcı (BMP) ortak etkisine (%32,1) bağlı olduğu görülmektedir. Bu, ölçülemeyen değişkenlik ve hata kaynaklarının fazla olduğuna işaret etmektedir. Ortak etki varyansından sonraki en büyük varyans kaynağı %30,6 ile birey-madde (BM) etkileşiminden kaynaklanmaktadır. Bu veri, belirli bireylerin bir görevden diğerine bağıl durumlarındaki farklılıklarının arttığı şeklinde yorumlanabilir. Toplam varyans içindeki en düşük paya sahip olan puanlayıcı (P) yüzeyine atfedilebilen varyans yüzdesi (%0), öğretmenlerin gerçekleştirdiği puanlamaların genel olarak önemli ölçüde değişmediğini, katılık veya hoşgörü göstermediğine işaret etmektedir. Birey ve puanlayıcı (BP) arasındaki etkileşimin,

gözlemlenen varyansın %4,4'ünü açıkladığı göz önüne alındığında, öğretmenlerin belirli öğrencilere yönelik puanlamalarının bir miktar değişmekle birlikte bu değişimin nispeten düşük olduğu yorumu yapılabilir. Madde ve puanlayıcı (MP) arasındaki etkileşim, toplam varyansa %3,7 civarında katkıda bulunmuştur. Bu varyans yüzdesinin küçük olması, maddeler arasındaki puanlama davranışında çok küçük değişiklikler olduğunu göstermektedir. Ölçme objesi olan bireylerin toplam varyans içerisindeki oranı üçüncü en büyük varyans oranı olan %20,4'tür. Bireylerden kaynaklanan varyans oranının büyük olması genelde istenen bir durumdur. Bu puanlayıcıların öğrencilerin arasındaki farklılıkları ortalama düzeyde ayırabildiğini gösterir. Son olarak, madde yüzeyinin (M) toplam varyans içindeki payının düşük olduğu söylenebilir. (%9,1). Bu, maddelerin bir maddeden diğerine zorluk ve ayırt edicilik olarak değişkenlik göstermediği şeklinde yorumlanabilir.

Tablo 20'daki deneyimli puanlayıcılara ait sütun incelendiğinde, Deneyimsiz puanlayıcılarda olduğu gibi toplam varyansın %0'ı puanlayıcılardan (P) kaynaklanan etkilerle ilişkilidir. Bu durum göz önünde bulundurulduğunda direkt olarak puanlayıcılardan kaynaklı bir değişkenliğin bulunmadığı söylenebilir. Bireylerden kaynaklanan varyans (B), %17,5 civarındadır ve toplam varyans içinde en büyük üçüncü paya sahiptir. Bireyler için kestirilen varyans değeri payının büyüklüğü, bireyler arasındaki sistematik farklılıkların ortaya konulduğunu göstermektedir. Bu bağlamda her ne kadar en büyük pay bu varyans bileşenine ait olmasa da belirli bir oranda birey farklılıklarının ayrıldığı söylenebilir. Görevlerden (M) kaynaklanan değişkenlik etkisi %11,1 ile nispeten orta düzeydedir. Bu oran doğrultusunda görev güçlük düzeylerinin benzeşik olmaktan ziyade küçük bir miktar da olsa farklılığa sahip olduğu söylenebilir. Birey-madde ortak etkisi (BM) incelendiğinde, toplam varyans içindeki en büyük payın %38,2 ile bu etkiden kaynaklandığı söylenebilir. Bu durum, belirli bireylerin bir görevden diğerine bağıl durumlarındaki farklılıklarının arttığı şeklinde yorumlanabilir. Birey-puanlayıcı (BP, %3,4) ve madde-puanlayıcı (MP, %0,5) ortak etkileri incelendiğinde atfedilen varyans oranlarının nispeten düşük olduğu görülmektedir. Bu durum, belirli puanlayıcıların belirli bireyler arasındaki puanlamalarının çok az değişiklik gösterdiği ve puanlayıcıların görevler arasında nispeten kararlı puanlamalar yaptıkları şeklinde yorumlanabilir. Son olarak, en büyük ikinci varyans yüzdesinin artık hata ile birlikte anılan birey, madde

ve puanlayıcı (BMP) ortak etkisine (%29,3) bağlı olduğu görülmektedir. Bu, ölçülemeyen değişkenlik kaynaklarının çokluğuna işaret etmektedir.

Deneyimsiz puanlayıcılara yönelik yapılan G çalışması sonucu ölçme objesine bağlı varyans bileşenlerinin toplamı sonucu elde edilen bağıl hata varyansı 0,09319, G katsayısı $E\rho^2$=0,83 ve Phi katsayısı Φ=0,81 olarak kestirilmiştir. Kestirilen G ve Phi katsayılarının (≥ 0,80) istenen düzeyde olduğu görülmektedir. Deneyimsiz 9. Sınıf Türk dili ve edebiyatı öğretmenleri, Yazma Becerileri Testinde, G kuramına göre, aşırı katı veya hoşgörülü olmadan ve bireyler ile görevler arasında değişiklik göstermeden puanlama gerçekleştirmişlerdir. Bununla birlikte kaynağı belli olmayan etkilerin nispeten fazla olduğu görülmektedir. Deneyimli puanlayıcılara yönelik gerçekleştirilen G çalışması sonucu, bağıl hata varyansı 0,07828, G katsayısı $E\rho^2$=0,81 ve Phi katsayısı da Φ=0,80 olarak bulunmuştur. Kestirilen G ve Phi değerlerinin istenilen seviyede (≥ 0,80) olduğu görülmektedir. Bununla birlikte, G kuramı altında yapılan analiz sonuçlarına göre, puanlama deneyimi olmayan Türkçe öğretmenlerinin yaptığı puanlamalara yönelik G katsayısı (0,83) ve Phi katsayısı (0,81), puanlama deneyimi olan Türkçe öğretmenlerine ait G (0,81) ve Phi (0,80) katsayılarına göre daha yüksek bulunmuştur.

Bölüm 5
Sonuç, Tartışma ve Öneriler

Sonuçlar

Bu araştırmada 2017 yılında Ankara, Adana ve İstanbul illerinde bulunan 4, 7 ve 9. Sınıf öğrencilerine uygulanan yazılı anlatım becerileri testi öğrenci yanıtlarına yönelik elde edilen verileri puanlayan deneyimli ve deneyimsiz puanlayıcılara ait güvenirlik indeksleri ve uyum düzeyleri, KTK'ya dayalı Kappa istatistiği ve Krippendorff'un Alpha tekniği ile Genellenebilirlik kuramına dayalı yöntemler kullanılarak belirlenmeye çalışılmıştır.

Gerçekleştirilen analizler sonucunda Kappa istatistiği ve Krippendorff'un Alpha tekniğinden elde edilen sonuçlar birbirleriyle paralellik göstermiştir. G kuramına ait sonuçlar incelendiğinde ise değişkenlik kaynağını göstermesi sebebiyle bu yöntem ile daha fazla bilgiye ulaşıldığı yorumu yapılabilir.

Sınıf öğretmenlerine yönelik puanlayıcı güvenirliği sonuçları.

- KTK'ya dayalı Kappa ve Krippendorff'un Alpha istatistikleri incelendiğinde puanlama deneyimi olmayan sınıf öğretmenleri; birinci, ikinci ve üçüncü cümle yazma görevlerinde (C1, C2, C3) puanlama deneyimine sahip sınıf öğretmenlerine göre nispeten daha yüksek uyum düzeyi göstermişlerdir. Dördüncü cümle görevinde (C4) ise deneyimli sınıf öğretmenlerinin puanlayıcı uyumu daha yüksektir.
- Paragraf yazma görevlerinde birinci ve üçüncü paragraf görevlerinin tüm ölçütlerinde (P1Ö1, P1Ö2, P1Ö3, P1Ö4, P3Ö1, P3Ö2, P3Ö3, P3Ö4) puanlama deneyimi olmayan sınıf öğretmenlerine ait uyum düzeyleri daha yüksek iken puanlama deneyimine sahip sınıf öğretmenleri ikinci paragraf görevinin tüm ölçütlerinde (P2Ö1, P2Ö2, P2Ö3, P2Ö4) daha yüksek uyum düzeyi yakalamışlardır.
- Metin görevi incelendiğinde ise tüm ölçütlerde (M1Ö1, M1Ö2, M1Ö3, M1Ö4, M1Ö5, M1Ö6, M1Ö7, M1Ö8) puanlama deneyimi olmayan sınıf öğretmenlerinin uyum düzeyleri daha yüksek bulunmuştur.
- G kuramına yönelik analizler incelendiğinde $E\rho^2$ ve Φ değerlerine göre puanlama deneyimlerine sahip sınıf öğretmenlerinin puanlayıcı

güvenirlik düzeylerinin deneyimsiz gruba göre daha yüksek olduğu gözlenmiştir ($E\rho^2 = 0,82$ ve $\Phi = 0,79$). Hata varyansının (BMP) %31 gibi nispeten yüksek oranda olması her iki grup için de kaynağı belli olmayan etkilerin yüksek olduğu şeklinde yorumlanabilir. Her iki grupta da puanlayıcıya atfedilen değişkenlik kaynağına yönelik varyans %0, puanlayıcının dâhil olduğu diğer değişkenlik kaynaklarının nispeten düşük olduğu sonucuna ulaşılmıştır. Bununla birlikte her iki puanlayıcı grubunda en yüksek varyans kaynağı ölçülemeyen varyans olarak nitelendirilen birey madde puanlayıcı (BMP) ortak etkisine atfedilmiştir.

- İstenilen düzeyde Phi değerine (0,80) ulaşılamadığından gerçekleştirilen D (karar) çalışmaları sonucunda puanlama deneyimi olmayan sınıf öğretmenleri için diğer yüzeyler sabit tutulduğunda en uygun puanlayıcı sayısı 12 ($\Phi = 0,80$) ve puanlama deneyimine sahip sınıf öğretmenleri için de en uygun puanlayıcı sayısı 4 ($\Phi = 0,84$) olarak kestirilmiştir.

Türkçe öğretmenlerine yönelik puanlayıcı güvenirliği sonuçları.

- KTK'ya dayalı Kappa ve Krippendorff'un Alpha istatistikleri incelendiğinde puanlama deneyimi olmayan 7. sınıf Türkçe öğretmenlerinin tüm cümle yazma görevlerinde (C1, C2, C3, C4) deneyimli gruba göre daha düşük uyum düzeyine sahip olduğu görülebilir.
- Paragraf yazma görevlerinde birinci paragraf görevi ikinci ve üçüncü ölçütü, ikinci paragraf görevi birinci, ikinci ve üçüncü ölçütleri ve üçüncü paragraf görevindeki tüm ölçütlerde (P1Ö2, P1Ö3, P2Ö1, P2Ö2, P2Ö3, P3Ö1, P3Ö2, P3Ö3, P3Ö4, P3Ö5) deneyimsiz sınıf öğretmenleri deneyimli gruba göre daha yüksek uyum sergilemişlerdir. Geriye kalan paragraf görevlerine ait ölçütlerde (P1Ö1,P1Ö4, P2Ö4) deneyimli grubun uyum düzeyleri daha yüksektir.
- Metin görevine ait ölçütler incelendiğinde birinci, ikinci, üçüncü, dördüncü, beşinci ve sekizinci ölçütlerde (M1Ö1, M1Ö2, M1Ö3, M1Ö4, M1Ö5, M1Ö8) deneyimli grubun uyum düzeyleri daha yüksek iken geriye kalan altıncı ve yedinci ölçütlerde (M1Ö6, M1Ö7)

deneyimsiz grubun uyum dizeylerinin daha yüksek olduğu gözlenmiştir.

- G kuramına yönelik analizler incelendiğinde $E\rho^2$ ve Φ değerlerine göre tüm ölçek üzerinde puanlama deneyimi olmayan 7. sınıf Türkçe öğretmenlerinin puanlayıcı güvenirlik düzeylerinin deneyimli gruba göre daha yüksek olduğu gözlenmiştir ($E\rho^2 = 0,87$ ve $\Phi = 0,86$). Deneyimli gruptaki puanlayıcılara atfedilen değişkenlik kaynağına yönelik varyans %8,1 iken deneyimsiz grupta bu varyans oranı %0,1 olarak bulunmuştur. Puanlayıcının dâhil olduğu diğer değişkenlik kaynaklarında da deneyimli grubun varyans kaynakları deneyimsiz gruba göre daha yüksek bir orana sahiptir. Bununla birlikte deneyimsiz grubun ölçülemeyen varyans olarak nitelendirilen birey madde puanlayıcı (BMP) ortak etkisi daha yüksek bulunmuştur. Bu değişkenlik kaynağının yüksek olması kaynağı belli olmayan değişkenlerin sürece karıştığı şeklinde yorumlanabilir.
- Deneyimli grupta istenilen düzeyde Phi değerine (,80) ulaşılamadığından gerçekleştirilen D (karar) çalışmaları sonucunda puanlama deneyimi olan Türkçe öğretmenleri için diğer yüzeyler sabit tutulduğunda en uygun puanlayıcı sayısı 3 ($\Phi =,80$) olarak kestirilmiştir.

Türk dili ve edebiyatı öğretmenlerine yönelik puanlayıcı güvenirliği sonuçları.

- KTK'ya dayalı Kappa ve Krippendorff'un Alpha istatistikleri incelendiğinde puanlama deneyimi olmayan 9. sınıf Türk dili ve edebiyatı öğretmenlerinin birinci, ikinci ve dördüncü cümle yazma görevlerinde (C1, C2, C4) deneyimli gruba göre daha yüksek uyum düzeyine sahip olduğu sonucuna ulaşılmıştır. Üçüncü cümle görevinde (C3) ise deneyimli grubun uyum düzeyi daha yüksektir.
- Paragraf yazma görevlerinde birinci ve ikinci paragraf görevlerine ait ölçütlerin tümünde (P1Ö1, P1Ö2, P1Ö3, P1Ö4, P2Ö1, P2Ö2, P2Ö3, P2Ö4) deneyimli grup deneyimsiz gruba göre daha yüksek uyum düzeyi yakalamıştır. Üçüncü paragraf görevinde ise sadece üçüncü

ölçüt (P3Ö3) için deneyimli grubun uyum düzeyi daha yüksek iken diğer ölçütlerde (P3Ö1, P3Ö2, P3Ö4, P3Ö5) deneyimsiz grubun uyum derecesi daha yüksek bulunmuştur.

- Metin görevine ait tüm ölçütlerdeki (M1Ö1, M1Ö2, M1Ö3, M1Ö4, M1Ö5, M1Ö6) puanlayıcı uyum düzeyleri ilgi çekici bir şekilde deneyimsiz grup lehine iken deneyimli grubun uyum düzeyleri nispeten düşük seviyede kalmıştır.
- G kuramına yönelik analizler incelendiğinde $E\rho^2$ ve Φ değerlerine göre tüm ölçek üzerinde puanlama deneyimi olmayan 9. sınıf Türk dili ve edebiyatı öğretmenlerinin güvenirlik düzeylerinin deneyimli gruba göre daha yüksek olduğu gözlenmiştir ($E\rho^2 = 0,83$ ve $\Phi = 0,81$). Burada gözden kaçırılmaması gereken bir nokta da deneyimli gruba ait Phi değerinin de $\Phi = 0,80$ nispeten yüksek olduğudur. Bu bağlamda iki grup arasındaki farkın makul seviyede düşük olduğu ve her iki gruptaki puanlayıcıların da güvenilir derecede puanlama yaptığı söylenebilir. Her iki grupta da puanlayıcılara atfedilen değişkenlik kaynağına yönelik varyans %0 bulunmuştur. Puanlayıcıların dâhil olduğu diğer değişkenlik kaynaklarında da deneyimli ve deneyimsiz grubun varyans kaynaklarının oranı tüm varyans içinde nispeten düşüktür. Her iki puanlayıcı grubunda ölçülemeyen varyans olan birey madde puanlayıcı (BMP) ortak etkisi toplam varyans içinde en yüksek paya sahiptir. Deneyimli ve deneyimsiz grupta istenilen düzeyde Phi değerine (0,80) ulaşıldığı için herhangi bir D (karar) çalışması yapılmasına ihtiyaç duyulmamıştır.
- G kuramı yöntemiyle gerçekleştirilen analizler sonucunda elde edilen güvenirlik değerleri tüm sürece yönelik bir atıfta bulunulabilmesine olanak sağlayabilir. Ancak değişkenlik kaynaklarındaki kaynağı belli olmayan hatalar oransal olarak fazla olduğu için güvenilir puanlama gerçekleştirildiğine dair bir çıkarım yapılması doğru olmayabilir. Değişkenlik kaynaklarının oransal olarak ölçme objesi olan bireylere, puanlayıcılara ve görevlere atfedilmesinin önemli olduğu

düşünüldüğünde bu çalışmadaki kaynağı belli olmayan hataların çokluğu güvenirliğin nispeten ektilendiği anlamına gelebilir.

Tartışma

Bilişsel özelliklerin ölçülmesinde, özellikle kesin doğru cevabın tam olarak kestirilmesinin zor olduğu açık uçlu maddelerdeki puanlayıcı etkilerini belirlemeye yönelik kullanılabilen G kuramı çalışmaları değişkenlik kaynaklarını adres göstermesiyle önemli bir teori olarak eğitim bilimcilerin karşısına çıkabilmektedir. Açık uçlu maddeler değerlendirilirken puanlayıcı, belirli öğrencileri puanlarken önyargılıysa, bu bireyler için olumsuz sonuçlara yol açabilir. Bu, tek bir puanlayıcının kullanıldığı durumlarda ortaya çıkabilecek olası bir olumsuz sonuç olarak düşünülmelidir (Messick, 1995). Bu sebeple birden fazla puanlayıcının kullanıldığı durumlar daha güvenilir kabul edilmektedir. Deneyimli puanlayıcıların nispeten daha uyumlu sonuçlar elde etmesi ilk etapta akla gelebilecek iken bu çalışma sonucunda her iki deneyim grubunda da farklı oranlarda uyum düzeyleri gözlenmiştir.

Öneriler

Araştırma sonuçlarından hareketle uygulama ve gelecek araştırmalara ilişkin yapılabilecek önerilerden önemli olduğu düşünülenler aşağıda sıralanamıştır.

Puanlamanın güvenirliği, özellikle geniş ölçekli sınavlarda bireylerin gelecek eğitim ve hayatlarını etkileyen önemli bir husustur. Puanlama güvenirliğinin önemli bir özelliği de daha önce değinildiği gibi aslında puanlayıcılar arası güvenirliktir. Bu bakımdan puanlama kriterlerinin net olması ve nitelikli puanlama anahtarlarının geliştirilmesi puanlama sürecinde büyük önem arz etmektedir. Araştırmacılar puanlayıcılar arası güvenirlik bağlamında birçok çalışma yapmış ve çeşitli yöntemlerin gelişmesi söz konusu olmuştur. Bu güvenirlik kestirimi yapılırken güncel literatür üç yaklaşım sunmaktadır, bunlar güvenirliğin görüş birliğine dayandığı puanlayıcılar arası anlaşma/uyum, puanlayıcılar arası güvenirliğin ölçümü ve puanlayıcılar arası tutarlılık tahminleridir. Bu üç yöntemin de avantajları ve dezavantajları ve farklı ölçüm teknikleri ile farklı varsayımları mevcuttur (Rashid ve Mahmood, 2016). Daha eski ve köklü bir yaklaşım olan ve KTK'ya dayanan Kappa istatistiği ile Krippendorff'un Alphası hesaplanması nispeten kolay olsa da uyumsuzluk kaynağını göstermediği, sınırlı sayıda puanlayıcıya uygulanabilirliği ve

ölçme durumlarındaki madde sayısı gibi farklı durumları hesaba katmadığı için daha az bilgi sağladığı düşünülebilir. Bunun yanında G kuramının, sağladığı bilgi ve değişkenlik kaynaklarına atıfta bulunması açısından daha tercih edilebilir olduğu çıkarımı yapılabilir.

Genellenebilirlik kuramı daha zengin bilgi vermesine karşın farklı modeller kullanmak da mümkündür. Bu çerçevede çaprazlanmış ve yuvalanmış modeller tasarlanılarak elde edilen sonuçlar karşılaştırılabilir. Bunun yanında Kappa ve Krippendorff teknikleri arasında farklı puanlayıcı grupları için farklı sonuçlar (sınıf öğretmenlerinde benzer ve ortaokul Türkçe öğretmenlerinde farklı olmak üzere) sonuçar elde edilmiştir. Bu konuda daha fazla araştırma yapılmaya gereksinim olduğu söylenebilir.

Bu araştırmada sadece yazma görevleri üzerinde karşılaştırma yapılmaya çalışılmıştır. Daha farklı performans görevleri, ödev, proje vb. için de bu yöntemlerin karşılaştırılması önerilmektedir.

Bu araştırmada puanlayıcı güvenirliğini belirlemeye yarayan yalnızca 3 yöntem kullanılmıştır. Bu konuda kullanılabilecek çok sayıda yöntem (Rasch çok yüzeyli analiz, log liner modelleri vb.) bulunmaktadır ve araştırmacılar yazma becerileri için diğer teknikleri kullanacakları araştırmalar tasarlayabilirler.

Türkiye'de son yıllarda PISA, TIMSS ve PIRLS gibi sınavlarda örnekleri bolca görülen üst düzey düşünme becerilerini ölçen madde türlerine ve görevlere yönelim olduğu söylenebilir. Bu kapsamda MEB'in özkaynaklarıyla geliştirdiği ABİDE ve Yazılı Anlatım Becerilerinin Ölçülmesi ve Değerlendirilmesi çalışmaları örnek gösterilebilir. MEB ayrıca ABİDE'yi Türkiye'nin geniş ölçekli merkezi sınavlarında kullanılanlardan daha geniş bir soru yelpazesini denemek için kullanmaktadır. Bu kapsamda açık uçlu, yapılandırılmış yanıtlı maddelere bu sınavda yer verilmektedir (Kitchen, Bethell, Fordham, Henderson ve Li, 2019). Uzun vadede geniş ölçekli sınavlarda bu tarz sorulara yer verme hedefinin olduğu düşünüldüğünde puanlama süreçlerinin iyileştirilmesi ve nitelikli puanlama anahtarlarının geliştirilmesi çalışmalarına ağırlık verilmesinin faydalı olacağı düşünülebilir. Geniş ölçekli sınavların yanında çağın gerektirdiği üst düzey düşünme becerilerinin ölçülmesi anlamında yapılacak merkezi sınavlar da geniş kitlelere ulaşma bakımından önemlidir. Çalışma sonuçlarına göre, puanlama deneyimine sahip olmamasına

rağmen nispeten güvenilir puanlama gerçekleştiren öğretmenler olduğu sonucuna ulaşılabilir. Bu durum, görevlerin ve puanlama anahtarlarının niteliğinin puanlama sonuçlarını olumlu yönde etkileyebileceğini göstermektedir.

Uygulamaya yönelik öneriler.

- Yazılı anlatım becerilerinin değerlendirildiği çalışmada puanlayıcı eğitimleri ve seçimlerinde bazı kriterler oluşturulabilir ve bu doğrultuda düzenlemeler yapılabilir. Araştırma MEB tarafından gerçekleştirilen bir uygulamadan elde edilen veriler üzerinden yürütüldüğü için böyle bir düzenleme yapılması veya bu sürecin takibi mümkün olmamıştır.
- Puanlama süresince yapılacak anlık dönütlerle puanlayıcılara olumlu yönlendirmelerde bulunabilir. Bu sayede örneğin puanlama deneyimine sahip olmayan sınıf öğretmenlerinin ilk üç cümle görevinde nispeten yüksek puanlayıcı uyumu göstermesine rağmen dördüncü cümle görevinde çok düşük puanlama yapmasının nedenleri ortaya çıkarılabilir ve görevin yanlış anlaşılması, puanlama anahtarındaki bir eksiklik veya yönergelerdeki varsa aksaklıklara yerinde müdahale gibi düzenlemeler gerçekleştirilebilir.
- Puanlayıcıların birbirlerinden etkilenme durumları göz önünde bulundurularak sürecin daha bağımsız olması açısından geliştirmeler önemli olabilir. Bu araştırmada puanlama sürecini gözlemleme olanağı bulunmadığı için böyle bir kontrol süreci de mümkün olmamıştır.
- Mevcut çalışmanın farklı yöntemler ve teknikler kullanılarak geliştirilmesi ve devam etmesinin de önem arz ettiği düşünülebilir.
- Çalışma sonucunda yapılacak kapsamlı analizler ve alan uzmanı görüşleri doğrultusunda iyileştirmeler yapılabileceği düşünülmektedir.

Gelecek araştırmalara yönelik öneriler.

- Puanlama süreçlerinde KTK'ya dayalı modellerin yanında daha güncel ve bilgi sağlayıcı modellerin kullanılmasına önem gösterilebilir. Puanlama güvenirliğine ilişkin son gelişmeler ve yoğun ilgi nedeniyle birçok yeni model geliştirilmiştir. Çoktan seçmeli testlerin bilişsel özellikleri ölçmede belirli bir seviyenin üzerine çıkamaması açık uçlu

yapılandırılmış yanıtlı madde tiplerine olan ilginin artmasına neden olmuştur. Özellikle sınıf içi ölçme uygulamalarında öğretmenlerin sıklıkla başvurdukları yazılı yoklama tipi sınavların objektif değerlendirilmesi açısından bu alanda yapılacak çalışma ve araştırmalar söz konusu öğretmenlerin hizmet içi gibi eğitim programlarına yol gösterecek bulgulara ulaşılması açısından da büyük öneme sahip olduğu düşünülebilir.

Kaynaklar

AERA, APA, and NCME (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education). 1999. *Standards for Educational and Psychological Testing.* Washington, DC: American Psychological Association.

Aihara, K., Au, K., Carroll, J., Nakanishi, P., Scheu, J., & Wong-Kam, J. A. (1999). The Professional Library: The Reading-Writing Connection for Struggling Readers. *The Reading Teacher*, *53*(3), 206-208.

Akbaş, O., & Özdemir, S. M. (2002). Avrupa Birliğinde yaşam boyu öğrenme. *Milli Eğitim Dergisi*, *155*(156), 112-126.

AlKhoudary, Y. A. M. (2015). The effect of teaching critical thinking on Al-Buraimi University College students' writing skills: A case study. *International Journal of applied linguistics and English literature*, *4*(6), 212-219.

Anadol, H. Ö. (2017). *Dereceli Puanlama Anahtarlarının Güvenirliğinin Farklı Deneyim Yıllarına Sahip Puanlayıcıların Kullanıldığı Durumlarda İncelenmesi* (Yüksek Lisans Tezi). Ankara Üniversitesi, Ankara.

Arifin, W. N. (2018). A Web-based Sample Size Calculator for Reliability Studies. *Education in Medicine Journal*, *10*(3).

Atilgan, H. (2013). Sample size for estimation of g and phi coefficients in generalizability theory. *Eurasian Journal of Educational Research*, *51*, 215-227.

Attali, Y. (2016). A comparison of newly-trained and experienced raters on a standardized writing assessment. *Language Testing*, *33*(1), 99-115.

Bangert-Drowns, R. L., Hurley, M. M., & Wilkinson, B. (2004). The effects of school-based writing-to-learn interventions on academic achievement: A meta-analysis. *Review of educational research*, *74*(1), 29-58.

Barkaoui, K. (2008). *Effects of Scoring Method and Rater Experience on ESL Essay Rating Processes and Outcomes* (Doktora tezi). University of Toronto, Toronto.

Barkaoui, K. (2010). Do ESL essay raters' evaluation criteria change with experience? A mixed-methods, cross-sectional study. *Tesol Quarterly*, 44(1), 31-57.

Barnwell, D. (1989). 'Naive'native speakers and judgements of oral proficiency in Spanish. *Language Testing*, *6*(2), 152-163.

Barritt, L., Stock, P. L., & Clark, F. (1986). Researching practice: Evaluating assessment essays. *College Composition and Communication*, *37*(3), 315-327.

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

Bıkmaz, Ö. (2011). *Üst düzey zihinsel özelliklerin ölçülmesinde puanlayıcılar arası güvenirlik belirleme tekniklerinin karşılaştırılması* (Yüksek lisans tezi). Hacettepe Üniversitesi, Ankara.

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

Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (pp. 349–381). Jossey-Bass.

Block, C. C., & Pressley, M. (2007). Best Practices in Teaching Comprehension. In L. B. Gambrell, L. M. Morrow, & M. Pressley (Eds.), *Best practices in literacy instruction* (pp. 220–242). Guilford Press.

Bouwer, R., Béguin, A., Sanders, T., & Van den Bergh, H. (2015). Effect of genre on the generalizability of writing scores. *Language Testing*, *32*(1), 83-100.

Brennan, R. L. (1992). Generalizability theory. *Educational Measurement: Issues and Practice*, *11*(4), 27-34.

Brennan, R. L., & National Council on Measurement in Education. (2006). *Educational measurement*. Praeger Publishers.

Brennan, R.L. (2003). *Coefficients and Indices in Generalizability Theory*. CASMA

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

Brown, S. (1986). *Reading-Writing Connections: College Freshman Basic Writers' Apprehension and Achievement*. Paper presented at the Annual Meeting of the Conference on College Composition and Communication - New Orleans, LA, March 13-15, 1986.

Büyükkıdık, S. (2012). *Problem çözme becerisinin değerlendirilmesinde puanlayıcılar arası güvenirliğin klasik test kuramı ve genellenebilirlik kuramına göre karşılaştırılması* (Yüksek lisans tezi). Hacettepe Üniversitesi, Ankara.

Cantor, A. B. (1996). Sample-size calculations for Cohen's kappa. *Psychological methods*, *1*(2), 150.

Carballo-Fazanes, A., Rey, E., Valentini, N. C., Rodríguez-Fernández, J. E., Varela-Casal, C., Rico-Díaz, J., & Abelairas-Gómez, C. (2021). Intra-Rater (Live vs. Video Assessment) and Inter-Rater (Expert vs. Novice) Reliability of the Test of Gross Motor Development—Third Edition. *International Journal of Environmental Research and Public Health*, *18*(4), 1652.

Cardinet, J., Johnson, S., & Pini, G. (2011). *Applying generalizability theory using EduG*. New York, NY: Routledge/Taylor & Francis Group.

Carillo, E. (2016). Engaging sources through reading-writing connections across the disciplines. *Across the Disciplines*, *13*(1), 19.

Charney, D. (1984). The validity of using holistic scoring to evaluate writing: A critical overview. *Research in the Teaching of English*, 65-81.

Chmura Kraemer, H., Periyakoil, V. S., & Noda, A. (2002). Kappa coefficients in medical research. *Statistics in medicine*, *21*(14), 2109-2129.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological assessment*, *6*(4), 284.

Coffman, W. E., & Thorndike, R. L. (1971). *Educational measurement*.

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

Collins, J., & Lee, J. (2005). Bringing together reading and writing to enhance comprehension. In *Annual Meeting of the National Reading Conference, Miami, FL*.

Cone, J. D. (1977). The relevance of reliability and validity for behavioral assessment. *Behavior Therapy*, *8*(3), 411-426.

Cook, D. A., & Beckman, T. J. (2006). Current concepts in validity and reliability for psychometric instruments: theory and application. *The American journal of medicine*, *119*(2), 166-e7.

Cronbach, L. J., Gleser, G., Nanda, H. & Rajaratnam, N. (1972). The dependability of behavioral measurements. *Theory of generalizability for scores and profiles*. Wiley, New York.

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

Darroch, J. N., & McCloud, P. I. (1986). Category distinguishability and observer agreement. *Australian Journal of Statistics*, *28*(3), 371-388.

Dennis, J. L., & Swinth, Y. (2001). Pencil grasp and children's handwriting legibility during different-length writing tasks. *American Journal of Occupational Therapy*, *55*(2), 175-183.

Derkuş, E. (2009). *Puanlayıcılar arasındaki uzlaşmanın farklı tekniklerle incelenmesi* (Yüksek lisans tezi). Mersin Üniversitesi, Mersin.

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied measurement in education*, *4*(4), 289-303.

Ediger, M. (2002). *Reading for Enjoyment and Pleasure.* Opinion papers (Eric Document Reproduction Service No. ED 471 184).

Erdosy, M. U. (2003). Exploring variability in judging writing ability in a second language: A study of four experienced raters of ESL compositions. *ETS Research Report Series*, *2003*(1), i-62.

Etemadzadeh, A., Seifi, S., & Far, H. R. (2013). The role of questioning technique in developing thinking skills: The ongoing effect on writing skill. *Procedia-Social and Behavioral Sciences*, *70*, 1024-1031.

Faragher, L., & Huijser, H. (2014). Exploring evidence of higher order thinking skills in the writing of first year undergraduates. *The International Journal of the First Year in Higher Education*, *5*(2), 33-44.

Farahzad, F., & Emam, A. (2010). Reading-writing connections in EAP courses: Implications and applications. *Journal of Language Teaching & Research*, *1*(5).

Ferah, A. (2007). *Türkçe İlk Okuma-Yazmayı Öğrenme.* Ankara: Nobel yayınları.

Fleenor, J. W., Fleenor, J. B., & Grossnickle, W. F. (1996). Interrater reliability and agreement of performance ratings: A methodological comparison. *Journal of Business and Psychology*, *10*(3), 367-380.

Freedman, S. W. (1979). How characteristics of student essays influence teachers' evaluations. *Journal of Educational Psychology*, *71*(3), 328.

Gambrell, L. B., & Mazzoni, S. A. (1999). Principles of best practice: Finding the common ground. *Best practices in literacy instruction*, 11-21.

Ganapathy, M., & Kaur, S. (2014). ESL Students' Perceptions of the Use of Higher Order Thinking Skills in English Language Writing. *Advances in Language and Literary Studies*, *5*(5), 80-87.

Gao, X., Brennan, R., & Guo, F. (2015). *Modeling measurement facets and assessing generalizability in a large-scale writing assessment*. GMAC Research Report RR 15-01.

Graham, M., Milanowski, A., & Miller, J. (2012). Measuring and Promoting Inter-Rater Agreement of Teacher and Principal Performance Ratings. *Online Submission*.

Gronlund, N. E. (1998). *Assessment of student achievement*. Allyn & Bacon Publishing, Longwood Division, 160 Gould Street, Needham Heights, MA 02194-2310.

Güler, N., & Teker, G. T. (2015). Açık uçlu maddelerde farklı yaklaşımlarla elde edilen puanlayıcılar arası güvenirliğin değerlendirilmesi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *6*(1), 12-24.

Gupta, T., Burke, K. A., Mehta, A., & Greenbowe, T. J. (2015). Impact of guided-inquiry-based instruction with a writing and reflection emphasis on chemistry students' critical thinking abilities. *Journal of Chemical Education*, *92*(1), 32-38.

Gürlen, E, Boztunç Öztürk, N, Eminoğlu, E. (2019). Investigation of the Reliability of Teachers, Self and Peer Assessments at Primary School Level with Generalizability Theory. *Journal of Measurement and Evaluation in Education and Psychology*, 10 (4) , 406-421.

Gwet, K. (2002). Kappa statistic is not satisfactory for assessing the extent of agreement between raters. *Statistical methods for inter-rater reliability assessment*, *1*(6), 1-6.

Gwet, K. L. (2011). *On the Krippendorff's alpha coefficient*. Gaithersburg, MD: Advanced Analytics.

Gwet, K. L. (2014). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC.

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in quantitative methods for psychology*, *8*(1), 23.

Hamp-Lyons, L. (1990). Second language writing: Assessment issues. *Second language writing: Research insights for the classroom*, 69-87.

Hamstra-Bletz, L., & Blöte, A. W. (1993). A longitudinal study on dysgraphic handwriting in primary school. *Journal of Learning Disabilities*, *26*(10), 689-699.

Han, T. (2013). *The impact of rating methods and rater training on the variability and reliability of efl students'classroom-based writing assessments in Turkish*

universities: An Investigation of Problems And Solutions (Doktora tezi). Atatürk Üniversitesi, Erzurum.

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, *1*(1), 77-89.

Huebner, A., & Skar, G. B. (2021). Conditional Standard Error of Measurement: Classical Test Theory, Generalizability Theory and Many-Facet Rasch Measurement with Applications to Writing Assessment. *Practical Assessment, Research & Evaluation*, 26(14).

Johnson, M. J. (2017). The rise of the citizen author: Writing within social media. *Publishing Research Quarterly*, *33*(2), 131-146.

Kan, A. (2001). Yazılı yoklamaların puanlanmasında puanlama cetveli ve yanıt anahtarı kullanımının puanlamaya ve puanlayıcı güvenirliğine etkisi. *Yayınlanmamış yüksek lisans tezi, Hacettepe Üniversitesi, Ankara*.

Kara, Y., & Kelecioğlu, H. (2015). Puanlayıcı niteliklerinin kesme puanlarının belirlenmesine etkisinin genellenebilirlik kuramı'yla incelenmesi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *6*(1), 58-71.

King, F.J.; Goodson, L.; Rohani, F. (2009*). Higher order thinking skills*. Center for Advancement of Learning and Assessment.

Kitchen, H, Bethell, G. Fordham, E. Henderson, K. and Li, R. R. (2019), OECD *Reviews of Evaluation and Assessment in Education: Student Assessment in Turkey*, OECD Publishing

Klein, S.P., Stecher, B.M., Shavelson, R.J., McCaffrey, D., Ormseth, T., Bell, R.M., Comfort, K. and Othman, A.R. (1998). Analytic versus holistic scoring of science performance tasks. *Applied Measurement in Education*, *11*(2), 121-137.

Klimova, B. F. (2013). Developing thinking skills in the course of academic writing. *Procedia-Social and Behavioral Sciences*, *93*, 508-511.

Knoch, U., Read, J., & von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing writing*, *12*(1), 26-43.

Knoeller, C. (2003). Imaginative response: Teaching literature through creative writing. *The English Journal*, *92*(5), 42-48.

Kozlowski, S. W., & Hattrup, K. (1992). A disagreement about within-group agreement: Disentangling issues of consistency versus consensus. *Journal of applied psychology*, *77*(2), 161.

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

Krippendorff, K. (2008). Systematic and random disagreement and the reliability of nominal data. *Communication Methods and Measures*, *2*(4), 323-338.

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology*. Sage publications.

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

Landy, F. J., Farr, J. F., & Farr, J. L. (1983). *The measurement of work performance: Methods, theory, and applications*. Academic Press.

Leary, M. R. (2014). *Introduction to behavioral research methods*. Pearson, 2014..

LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational research methods*, *11*(4), 815-852.

Leggette, H. R., McKim, B., Homeyer, M., & Rutherford, T. (2015). Perspectives of writing related to critical thinking and knowledge creation. *NACTA Journal*, *59*(4), 275.

Lim, G. S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing*, *28*(4), 543-560.

Lord, F. M., Novick, M. R., & Birnbaum, A. (1968). *Statistical theories of mental test scores.* Information Age Publications

Manalo, E., & Sheppard, C. (2016). How might language affect critical thinking performance?. Thinking Skills and Creativity, 21, 41-49.

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

Manfra, L., Squires, C., Dinehart, L. H., Bleiker, C., Hartman, S. C., & Winsler, A. (2017). Preschool writing and prematheматics predict grade 3 achievement for low-income, ethnically diverse children. *The Journal of Educational Research*, *110*(5), 528-537.

McColly, W. (1970). What does educational research say about the judging of writing ability? *The Journal of Educational Research*, *64*(4), 147-156.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American psychologist*, *50*(9), 741.

Milli Eğitim Bakanlığı. (2017). *Yazılı anlatım becerilerinin ölçülmesi ve değerlendirilmesi çalışması.* Ankara: Millî Eğitim.

Mitchell, S. K. (1979). Interobserver agreement, reliability, and generalizability of data collected in observational studies. *Psychological Bulletin*, *86*(2), 376.

Nitko, A. J., & Brookhart, S. M. (2016*). Öğrencilerin Eğitsel Değerlendirmesi.* Çeviri Editörleri: Bıçak, B. Bahar, M. ve Özel, S., Ankara: Nobel Yayıncılık.

Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of mathematical psychology, 3(1)*, 1-18.

Özdemir, E. (1987). Türkçe Öğretimi Kılavuzu (3. Basım). *İstanbul: İnkılâp Kitabevi*.

Paesani, K. (2016). Investigating connections among reading, writing, and language development: A multiliteracies perspective. *Reading in a Foreign Language*, *28*(2), 266-289.

Pantaleo, S. (2017). Critical thinking and young children's exploration of picturebook artwork. *Language and Education*, *31*(2), 152-168.

Parodi, G. (2007). Reading–writing connections: Discourse-oriented research. *Reading and writing*, *20*(3), 225-250.

Peltzer, K., & Machleidt, W. (1992). A traditional (African) approach towards the therapy of schizophrenia and its comparison with western models. *Therapeutic Communities*.

Popham, W. J. (1997). What's Wrong--and What's Right--with Rubrics. *Educational leadership*, *55*(2), 72-75.

Preiss, D. D., Castillo, J. C., Grigorenko, E. L., & Manzi, J. (2013). Argumentative writing and academic achievement: A longitudinal study. *Learning and Individual Differences*, *28*, 204-211.

Prior, P. A. (1995). Redefining the task: An ethnographic examination of writing and response in graduate seminars. In *Academic writing in a second language: Essays on research and pedagogy* (pp. 47-82). Norwood, NJ: Ablex.

Rashid, S., & Mahmood, N. (2016). High-Stake Testing in Punjab: Inter-rater Reliability in the Scoring of Secondary School Certificate (SSC) Examination. *Journal of Research & Reflections in Education (JRRE)*, *10*(2).

Reddy, Y. M., & Andrade, H. (2010). A review of rubric use in higher education. *Assessment & evaluation in higher education*, *35*(4), 435-448.

Reis, H. T., Gable, S. L., & Maniaci, M. R. (2014). *Methods for studying everyday experience in its natural context. In H. T. Reis & C. M. Judd* (Ed.), *Handbook of research methods in social and personality psychology* (373–403). Cambridge University Press

Şahan, Ö. (2018). *The impact of rater experience and essay quality on rater behavior and scoring* (Doktora tezi). Çanakkale Onsekiz Mart Üniversitesi, Çanakkale.

Şahan, Ö., & Razı, S. (2020). Do experience and text quality matter for raters' decision-making behaviors? *Language Testing*, *37*(3), 311-332.

Schoonen, R., Vergeer, M., & Eiting, M. (1997). The assessment of writing ability: Expert readers versus lay readers. *Language Testing*, *14*(2), 157-184.

Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. *American Psychologist*, *44*(6), 922-932.

Shohamy, E., Gordon, C. M., & Kraemer, R. (1992). The effect of raters' background and training on the reliability of direct writing tests. *The Modern Language Journal*, *76*(1), 27-33.

Sim, J., & Wright, C. C. (2005). The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical therapy*, *85*(3), 257-268.

Sireci, S. G. (1998). The construct of content validity. *Social indicators research*, *45*(1), 83-117.

Spearman, C. (1904). The proof and measurement of association between two things. *The American journal of psychology*, *15(1), 72–101*

Stemler, S. E., & Tsai, J. (2008). Best practices in interrater reliability: Three common approaches. *Best practices in quantitative methods*, 29-49.

Strube, M. J. (2002). *Reliability and generalizability theory*. In L.G. grimm & P.R. Yarnold (Eds.), *Reading and understanding more multivariate statistics* (pp. 23-66). Washington, DC: American Psychological Association.

Swiss Society for Research in Education Working Group. (2006). EduG (Version 6.1). https://www.irdp.ch/institut/english-program-1968.html kaynağından 21 Mart 2021 tarihinde alınmıştır.

Thorndike, R. M. ve Thorndike-Christ, T. (2017). *Psikolojide ve eğitimde ölçme ve değerlendirme* (M. Otrar, Çev. Ed.). Ankara: Nobel.

Tierney, R. J., & Shanahan, T. (1991). *Research on the reading–writing relationship: Interactions, transactions, and outcomes*. Lawrence Erlbaum Associates, Inc.

Tinsley, H. E., & Weiss, D. J. (1975). Interrater reliability and agreement of subjective judgments. *Journal of Counseling Psychology*, *22*(4), 358.

Tseng, M. H., & Cermak, S. A. (1993). The influence of ergonomic factors and perceptual–motor abilities on handwriting performance. *American Journal of Occupational Therapy*, *47*(10), 919-926.

Tunks, K. W. (2011). Exploring journals as a genre for making reading-writing connections. *Childhood Education*, *87*(3), 169-176.

Turgut, M. F., & Baykul, Y. (2012). *Eğitimde ölçme ve değerlendirme*. Pegem Akademi.

Ulrich, D. A., Ulrich, B. D., & Branta, C. F. (1988). Developmental gross motor skill ratings: A generalizability analysis. *Research Quarterly for Exercise and Sport*, *59*(3), 203-209.

Vach, W. (2005). The dependence of Cohen's kappa on the prevalence does not matter. *Journal of clinical epidemiology*, *58*(7), 655-661.

Van Der Linden, W. J., & Hambleton, R. K. (1997). Item response theory: Brief history, common models, and extensions. In *Handbook of modern item response theory* (pp. 1-28). Springer, New York, NY.

Waring, J. W. (2007). *The impact of writing on student achievement* (Doctoral dissertation, University of North Carolina Wilmington).

Webb, N. M., Rowley, G. L., & Shavelson, R. J. (1988). Using generalizability theory in counseling and development. *Measurement and Evaluation in Counseling and Development*, *21*(2), 81-90.

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing*, *11*(2), 197-223.

Weigle, S. C. (2002). *Assessing writing*. Ernst Klett Sprachen.

White, E. M. (1984). Holisticism. *College Composition and Communication*, *35*(4), 400-409.

Wolfe, E. W., & Feltovich, B. (1994). Learning to rate essays: A study of scorer cognition. April, *Paper presented at the annual meeting of the American Educational Research Association.*

Wright, D., & Piersel, W. C. (1992). Components of variance in behavior ratings from parents and teachers. *Journal of psychoeducational Assessment*, *10*(4), 310-318.

Yelboğa,A.(2008). Güvenirliğin Değerlendirilmesinde Genellenebilirlik Kuramı'nın Kullanılması: Endüstri ve Örgüt Psikolojisinde Bir Uygulama. *Psikoloji Çalışmaları Dergisi.* 28, 35-54.

Yıldıztekin, B. (2014). *Klasik test kuramı ve genellenebilirlik kuramından puanlayıcılar arası tutarlılığın farklı yöntemlere göre karşılaştırılması* (Yüksek lisans tezi). Hacettepe Üniversitesi, Ankara.

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

Zemelman, S., H. Daniels, and A. Hyde. 1998. *Best practice: New standards for teaching and learning in America's schools.* 2nd. ed. Portsmouth, NH: Heinemann

T.C.
MİLLÎ EĞİTİM BAKANLIĞI
Ölçme, Değerlendirme ve Sınav Hizmetleri
Genel Müdürlüğü

Sayı : 42497731-605.01-E.15445652 28.08.2019
Konu : Veri Talebi

Sayın: Mehmet Ali AYDOĞMUŞ

İlgi : 26.08.2019 tarihli ve 15303651 sayılı dilekçeniz.

İlgi dilekçeniz incelenmiş olup Daire Başkanlığımızdan talep etmiş olduğunuz veriler Cd' ile tarafınıza elden teslim edilecektir.

Bilgilerinize rica ederim.

Mustafa GELEN
Daire Başkanı

Adres: Emniyet Mh. Abant 2 Sk. 13/A Ek Bina Y.Mh. ANK.
Elektronik Ağ:odsgm.meb.gov.tr
e-posta:odsgm_argeprojelerdb@meb.gov.tr - Kep: meb@hs01.kep.tr

Ayrıntılı bilgi için: Fatih TÜRKMEN Bil. İşlt.
Tel: 0312 413 32 03
Faks: 0312 213 01 47

EK-C: Etik Beyanı

Hacettepe Üniversitesi Eğitim Bilimleri Enstitüsü, tez yazım kurallarına uygun olarak hazırladığım bu tez çalışmasında,

- tez içindeki bütün bilgi ve belgeleri akademik kurallar çerçevesinde elde ettiğimi,
- görsel, işitsel ve yazılı bütün bilgi ve sonuçları bilimsel ahlak kurallarına uygun olarak sunduğumu,
- başkalarının eserlerinden yararlanılması durumunda ilgili eserlere bilimsel normlara uygun olarak atıfta bulunduğumu,
- atıfta bulunduğum eserlerin bütününü kaynak olarak gösterdiğimi,
- kullanılan verilerde herhangi bir tahrifat yapmadığımı,
- bu tezin herhangi bir bölümünü bu üniversitede veya başka bir üniversitede başka bir tez çalışması olarak sunmadığımı

beyan ederim.

12/08/2021
Mehmet Ali
AYDOĞMUŞ

EK-Ç: Yüksek Lisans/Doktora Tez Çalışması Orijinallik Raporu

16/08/2021

HACETTEPE ÜNİVERSİTESİ
Eğitim Bilimleri Enstitüsü
Eğitim Bilimleri Ana Bilim Dalı Başkanlığına,

Tez Başlığı: Puanlayıcı Deneyimlerine Göre Puanlayıcılar Arası Güvenirliklerin Farklı Yöntemlerle İncelenmesi

Yukarıda başlığı verilen tez çalışmamın tamamı (kapak sayfası, özetler, ana bölümler, kaynakça) aşağıdaki filtreler kullanılarak **Turnitin** adlı intihal programı aracılığı ile kontrol edilmiştir. Kontrol sonucunda aşağıdaki veriler elde edilmiştir:

Rapor Tarihi	Sayfa Sayısı	Karakter Sayısı	Savunma Tarihi	Benzerlik Oranı	Gönderim Numarası
31/05/2021	82	22935	28/06 /2021	%9	1597826870

Uygulanan filtreler:
1. Kaynaklar hariç
2. Alıntılar dâhil
3. 5 kelimeden daha az örtüşme içeren metin kısımları hariç

Hacettepe Üniversitesi Eğitim Bilimleri Enstitüsü Tez Çalışması Orijinallik Raporu Alınması ve Kullanılması Uygulama Esasları'nı inceledim ve çalışmamın herhangi bir intihal içermediğini; aksinin tespit edileceği muhtemel durumda doğabilecek her türlü hukuki sorumluluğu kabul ettiğimi ve yukarıda vermiş olduğum bilgilerin doğru olduğunu beyan eder, gereğini saygılarımla arz ederim.

Ad Soyadı: Mehmet Ali AYDOĞMUŞ
Öğrenci No.: N17124593
Ana Bilim Dalı: Eğitim Bilimleri
Programı: Eğitimde Ölçme ve Değerlendirme
Statüsü: ☒ Y.Lisans ☐ Doktora ☐ Bütünleşik Dr.

İmza

DANIŞMAN ONAYI

UYGUNDUR.
(Prof. Dr. Nuri DOĞAN)

EK-D: Thesis/Dissertation Originality Report

16/08/2021

HACETTEPE UNIVERSITY
Graduate School of Educational Sciences
To The Department of Educational Sciences

Thesis Title: Analysis Of Interrater Reliability With Different Methods According to Rater Experience

The whole thesis that includes the *title page, introduction, main chapters, conclusions and bibliography section* is checked by using **Turnitin** plagiarism detection software take into the consideration requested filtering options. According to the originality report obtained data are as below.

Time Submitted	Page Count	Character Count	Date of Thesis Defense	Similarity Index	Submission ID
31/05/2021	82	22935	28/06 /2021	%9	1597826870

Filtering options applied:
1. Bibliography excluded
2. Quotes included
3. Match size up to 5 words excluded

I declare that I have carefully read Hacettepe University Graduate School of Educational Sciences Guidelines for Obtaining and Using Thesis Originality Reports; that according to the maximum similarity index values specified in the Guidelines, my thesis does not include any form of plagiarism; that in any future detection of possible infringement of the regulations I accept all legal responsibility; and that all the information I have provided is correct to the best of my knowledge.

I respectfully submit this for approval.

Name Lastname: Mehmet Ali AYDOĞMUŞ
Student No.: N17124593
Department: Educational Sciences
Program: Educational Measurement and Evaluation
Status: ☒ Masters ☐ Ph.D. ☐ Integrated Ph.D.

Signature

ADVISOR APPROVAL

APPROVED
(Prof. Dr. Nuri DOĞAN)

EK-E: Yayımlama ve Fikrî Mülkiyet Hakları Beyanı

Enstitü tarafından onaylanan lisansüstü tezimin/raporumun tamamını veya herhangi bir kısmını, basılı (kâğıt) ve elektronik formatta arşivleme ve aşağıda verilen koşullarla kullanıma açma iznini Hacettepe Üniversitesine verdiğimi bildiririm. Bu izinle Üniversiteye verilen kullanım hakları dışındaki tüm fikri mülkiyet haklarım bende kalacak, tezimin tamamının ya da bir bölümünün gelecekteki çalışmalarda (makale, kitap, lisans ve patent vb.) kullanım haklan bana ait olacaktır.

Tezin kendi orijinal çalışmam olduğunu, başkalarının haklarını ihlal etmediğimi ve tezimin tek yetkili sahibi olduğumu beyan ve taahhüt ederim. Tezimde yer alan telif hakkı bulunan ve sahiplerinden yazılı izin alınarak kullanılması zorunlu metinlerin yazılı izin alınarak kullandığımı ve istenildiğinde suretlerini Üniversiteye teslim etmeyi taahhüt ederim.

Yükseköğretim Kurulu tarafından yayınlanan **"Lisansüstü Tezlerin Elektronik Ortamda Toplanması, Düzenlenmesi ve Erişime Açılmasına ilişkin Yönerge"** kapsamında tezim aşağıda belirtilen koşullar haricince YÖK Ulusal Tez Merkezi / H.Ü. Kütüphaneleri Açık Erişim Sisteminde erişime açılır.

- o Enstitü/Fakülte yönetim kurulu kararı ile tezimin erişime açılması mezuniyet tarihinden itibaren 2 yıl ertelenmiştir. (1)
- o Enstitü/Fakülte yönetim kurulunun gerekçeli kararı ile tezimin erişime açılması mezuniyet tarihimden itibaren … ay ertelenmiştir. (2)
- o Tezimle ilgili gizlilik kararı verilmiştir. (3)

..... /..... /.....

Mehmet Ali AYDOĞMUŞ

"Lisansüstü Tezlerin Elektronik Ortamda Toplanması, Düzenlenmesi ve Erişime Açılmasına İlişkin Yönerge"

(1) Madde 6. 1. Lisansüstü tezle ilgili patent başvurusu yapılması veya patent alma sürecinin devam etmesi durumunda, tez danışmanının önerisi ve enstitü anabilim dalının uygun görüşü Üzerine enstitü veya fakülte yönetim kurulu iki yıl süre ile tezin erişime açılmasının ertelenmesine karar verebilir.

(2) Madde 6.2. Yeni teknik, materyal ve metotların kullanıldığı, henüz makaleye dönüşmemiş veya patent gibi yöntemlerle korunmamış ve internetten paylaşılması durumunda 3. şahıslara veya kurumlara haksız kazanç; imkânı oluşturabilecek bilgi ve bulguları içeren tezler hakkında tez danışmanın önerisi ve enstitü anabilim dalının uygun görüşü üzerine enstitü veya fakülte yönetim kurulunun gerekçeli kararı ile altı ayı aşmamak üzere tezin erişime açılması engellenebilir.

(3) Madde 7. 1. Ulusal çıkarları veya güvenliği ilgilendiren, emniyet, istihbarat, savunma ve güvenlik, sağlık vb. konulara ilişkin lisansüstü tezlerle ilgili gizlilik kararı, tezin yapıldığı kurum tarafından verilir. Kurum ve kuruluşlarla yapılan işbirliği protokolü çerçevesinde hazırlanan lisansüstü tezlere ilişkin gizlilik kararı ise, ilgili kurum ve kuruluşun önerisi ile enstitü veya fakültenin uygun görüşü Üzerine üniversite yönetim kurulu tarafından verilir. Gizlilik kararı verilen tezler Yükseköğretim Kuruluna bildirilir.*
Madde 7.2. Gizlilik kararı verilen tezler gizlilik süresince enstitü veya fakülte tarafından gizlilik kuralları çerçevesinde muhafaza edilir, gizlilik kararının kaldırılması halinde Tez Otomasyon Sistemine yüklenir

** Tez danışmanının önerisi ve enstitü anabilim dalının uygun görüşü üzerine enstitü veya fakülte yönetim kurulu tarafından karar verilir.*