Uluslararası Öğrenen Toplum Dergisi

e-ISSN: 3023-8374

2025 | Cilt 2 | Sayı 2

Sayfa 273-298

International Society That Learn Journal

e-ISSN: 3023-8374

2025 | Volume 2 | Issue 2

Page 273-298

Proje Yarışmalarında Farklı Puanlama Desenlerinin Güvenilirliği: Genellenebilirlik Kuramı ile Bir Simülasyon Çalışması

Reliability of Different Scoring Patterns in Project Competitions: A Simulation Study with Generalizability Theory

Mehmet IRMAK,

https://orcid.org/0009-0007-2298-8266

Esat Sivri Ortaokulu, Denizli, Türkiye,

irmak20@gmail.com

Yükleme: 19.09.2025; **Revizyon**: 11.11.2025; **Kabul:** 14.11.2025; **Yayınlanma:** 01.12.2025

Irmak, M. (2025). Proje Yarışmalarında Farklı Puanlama Desenlerinin Güvenilirliği: Genellenebilirlik Kuramı ile Bir Simülasyon Çalışması. *International Society that Learn Journal, 2*(2), 273-298.

Özet

Tüketen değil, üreten ve geliştiren nesillerin yetiştirilmesi adına, öğrencilerin bilgi çağının gerekliliği olan bilimsel araştırma yöntemlerini kullanarak öğrenmelerini sağlayan proje tabanlı öğrenme günümüzde önemli bir yere sahiptir. Bu bağlamda proje tabanlı öğrenmenin teşviki ve öğrencilerin bu alanda daha çok çalışma yapmalarına imkân sağlamak için ödüllü proje yarışmaları (TÜBİTAK, Teknofest, MEB vb.) düzenlenmektedir. Bu metodolojik araştırmanın amacı, proje yarışmalarına ait değerlendirme süreçlerinde genellenebilirlik kuramına dayalı farklı desenler oluşturarak farklı puanlayıcıların bu süreçler üzerindeki etkisini ortaya koymaktır. Çalışmada, robotik-kodlama ile ilgili proje yarışmalarının değerlendirme yapısını yansıtacak şekilde tasarlanmış simüle bir veri seti kullanılmıştır. 10 ayrı proje, proje yazımı konusunda uzman olan üç puanlayıcı ve bilişim teknolojileri alanında uzman altı puanlayıcı yer almıştır. Oluşturulan iki farklı desenin ilki olan çaprazlanmış desende (bxmxp) 10 proje, 3 proje uzmanı tarafından değerlendirilmiş, ikinci olarak yuvalanmış desende (bx(p:m)) puanlayıcılar maddelerde yuvalanarak 10 proje için bazı maddeleri proje yazım uzmanı, bazı maddeleri bilişim teknolojileri alan uzmanı puanlayıcılar puanlamıştır. Veriler analiz edilerek G ve K çalışmaları gerçekleştirilmiş, farklı madde ve puanlayıcı sayılarından elde edilen G ve Phi katsayıları ayrıntılı olarak incelenmiştir. Simülasyon sonuçlarına göre, çaprazlanmış ve yuvalanmış desenle yapılan ölçme işlemleri bağıl değerlendirmeler için yüksek güvenirlik sonuçları verirken, mutlak değerlendirme için uygun değerler görülmemiştir. Karar çalışmaları, puanlayıcı sayısı artırıldığında her iki desende de güvenirliğin arttığını göstermiştir. Bulgular, proje yarışmalarının değerlendirme süreçlerinin tasarlanmasında tüm puanlayıcıların tüm kriterleri değerlendirmesinin, uzmanlık alanlarına göre madde dağılımına kıyasla daha güvenilir sonuçlar üretebileceğini ortaya koymaktadır. Bu metodolojik karşılaştırma, eğitim kurumları ve organizatörlere değerlendirme süreçlerini tasarlarken kanıta dayalı kararlar alabilmeleri için önemli bilgiler sunmaktadır.

Anahtar Kelimeler: Genellenebilirlik kuramı, proje değerlendirme, puanlayıcı güvenirliği, simülasyon çalışması, çaprazlanmış desen, yuvalanmış desen

Abstract

Project-based learning, which enables students to learn by using scientific research methods required in the information age, holds an important place today in fostering generations who produce and innovate rather than merely consume. In this context, award-based project competitions (e.g., TÜBİTAK, Teknofest, Ministry of Education competitions) are organized to promote project-based learning and encourage students to engage more in this field. The aim of this methodological study is to construct different designs based on Generalizability Theory for the evaluation processes of project competitions and to reveal the influence of different raters on these processes. A simulated dataset reflecting the evaluation structure of project competitions on robotics and coding was used in the study. Ten separate projects, three raters specialized in project writing, and six raters specialized in information technologies participated. In the first of the two designs developed—the crossed design (b×m×p) 10 projects were evaluated by three project-writing experts; in the second, the nested design (b × (p:m)), raters were nested within items such that some items for each of the 10 projects were scored by project-writing experts while others were scored by information technology experts. G- and D-studies were conducted by analyzing the data, and G and Phi coefficients obtained under different numbers of items and raters were examined in detail. According to the simulation results, while measurement operations conducted with both the crossed and nested designs yielded high reliability for relative decisions, they did not produce adequate values for absolute decisions. Decision studies showed that reliability increased in both designs as the number of raters increased. The findings indicate that having all raters evaluate all criteria in project competitions may yield more reliable results than distributing items according to areas of expertise. This methodological comparison provides valuable insights to educational institutions and organizers for making evidence-based decisions when designing evaluation processes.

Keywords: Generalizability theory, project evaluation, rater reliability, simulation study, crossed design, nested design

Irmak, M. (2025). Proje Yarışmalarında Farklı Puanlama Desenlerinin Güvenilirliği: Genellenebilirlik Kuramı ile Bir Simülasyon Çalışması. *International Society that Learn Journal, 2*(2), 273-298.

Giriş

Günümüzde teknoloji, toplumları ekonomik, sosyal ve kültürel yönden destekleyerek gelişme ve büyüme süreçlerinde belirgin bir rol oynamaktadır. Özellikle teknolojik açıdan dünyaya egemen olan ülkeler, ileri teknolojiye sahip olarak global rekabet gücünü artırmakta ve dünya ekonomisinde her geçen gün daha da fazla söz sahibi olmaktadırlar. Bu nedenle teknolojiyi tüketen değil, üreten ve geliştiren bireylerin yetiştirilmesi, ülkelerin sürdürülebilir kalkınma hedeflerini gerçekleştirmelerinde kritik bir öneme sahiptir (Schwab, 2016). Bilgi çağının gereklilikleri de düşünüldüğünde, bu bireylerin bilim insanı perspektifiyle yaklaşabilen, problem çözme becerisine sahip, araştırmacı, sorgulama kimliğine sahip, üretken, yaratıcı düşünme becerilerine sahip bireyler olması gerekmektedir (Çetin ve Şengezer, 2013). Bu doğrultuda karşımıza çıkan proje tabanlı öğrenme yöntemlerinin kullanımı bireylerin teknolojiyi yalnızca tüketen değil, aynı zamanda üreten konumuna gelmeleri için önem arz etmektedir. Proje tabanlı öğrenme, eğitim-öğretim süreçlerinde gerçek hayatın içinden problemlerin ele alınarak çözüm yollarının ortaya konmasına yönelik bir yaklaşım sunmaktadır (Ayvaz Tuncel, 2021).

Eğitim-öğretim ortamlarında proje tabanlı öğrenmenin teşviki ve öğrencilerin bu alanda daha çok çalışmalarını özendirmek amacı ile ülkemizdeki birçok özel ve kamu kurumu farklı konu alanlarında ödüllü proje yarışmaları (TÜBİTAK, Teknofest, MEB vb.) düzenlemektedir. Günümüzde özellikle bilişim teknolojilerini kullanmayı teşvik eden proje yarışmaları daha çok ön plana çıkmakta, bu yarışmaların hazırlık sürecinde öğrenciler, farklı disiplinleri bir araya getirerek birçok farklı derste edindikleri teorik bilgileri, gerçek hayatta karşılaştıkları problemlerin çözümü üzerinde uygulama fırsatı bulmaktadırlar (TÜBİTAK, 2024). Ayrıca bu yarışmalar, katılımcı bireylerin bilimsel düşünme becerilerini geliştirmelerine, işbirlikçi takım çalışmaları yapmalarına ve yenilikçi fikirlerini hayata geçirmelerine olanak tanımaktadır (TÜBİTAK, 2019).

Bu yarışmalarda yürütülen süreçler incelendiğinde; bunların genellikle birden fazla değerlendirme aşaması içerdiği ve birçok puanlayıcının dereceleme puanlama anahtarı ile puan verdiği bir yapı üzerine kurulu olduğu görülmektedir (TÜBİTAK, 2024; MEB, 2024). Bu nedenle böylesine büyük yarışmalardaki en önemli bölüm, değerlendirme süreçleri olarak öne çıkmaktadır. Bu süreçlerin ve organizasyonu düzenleyen kurumların güvenirliği ve popülerliği düşünüldüğünde projeleri değerlendiren puanlayıcılar ve puanlama süreçlerinin güvenirliği önemli bir konu haline gelmektedir (Taştan ve Uzun, 2021). Ancak çok fazla proje başvurusunun gelmesi ve puanlayıcı sayısının kısıtlı olması nedeniyle öznel

e-ISSN: 3023-8374 © 2024 International Society that Learn Journal

değerlendirmenin ağırlıklı olduğu bu yarışmalarda jüri üyelerinin kararlarının genellenebilirliği sorgulanmakta ve bu kararların ne derece tutarlı olduğu belirsiz kalmaktadır (Brennan, 2001; Shavelson ve Webb, 1991). Puanlayıcılar, projelerin değerlendirilmesinde önemli bir değişkenlik ve hata kaynağıdır. Bununla birlikte, değerlendirme sürecine karışan diğer değişkenlik kaynakları ile araştırmada yer alan puanlayıcı, birey ve görev gibi unsurlar arasındaki etkileşimler de güvenirliği etkilemektedir (Güler, 2008). Bu nedenle, güvenilirlik değerlendirilirken, çeşitli hata kaynaklarının yanı sıra bu kaynakların birbirleriyle etkileşimlerinden kaynaklanabilecek hataların da göz önünde bulundurulması gerekmektedir (Taştan ve Uzun, 2021). Bu durumların üstesinden gelebilmek için çeşitli değişkenlerin etkisini analiz eden ve değerlendirme süreçlerinin daha sistematik bir şekilde yapılandırılmasını sağlayan genellenebilirlik kuramı uygulanabilir (Shavelson ve Webb, 1991). Özellikle ulusal ve uluslararası düzeyde yapılan teknoloji temelli proje yarışmaları gibi çok aşamalı ve birçok puanlayıcının dahil olduğu değerlendirme süreçlerinde, genellenebilirlik kuramına dayalı desenler, değerlendirmenin hem puanlayıcı değişkenliğini hem de puanlayıcı güvenilirliğini belirlemede güçlü bir yöntem olabilir (Brennan, 2001).

Genellenebilirlik kuramı (G kuramı), bir davranışın ölçülmesi sırasında güvenilirliğin değerlendirilmesini, güvenilir gözlem tasarımının yapılmasına, araştırma ve kavramlaştırmaya olanak sağlayan bir istatistik kuramıdır. Varyans analizi temel alınarak oluşturulan bu model, günümüzde hâlâ popüler olan klasik test kuramının gerçek puan modelindeki kısıtlamalarını aşmak amacıyla 1963 ve 1972 yıllarında Cronbach, Gleser, Nanda ve Rajaratnam tarafından geliştirilmiştir. Kullanım kolaylığı ve basit işlem adımları bu modelin tercih edilme sebeplerindendir (Atılgan, 2019). Genellenebilirlik çalışmalarında, ölçme güvenirliği için genellenebilirlik katsayısı (G-katsayısı) hesaplanır ve bu katsayı Cronbach alfa (α) katsayısı ile benzerlik gösterir (Price, 2021). Bu kuramın temel hedefi, mevcut örneklemden yola çıkarak genel evrene dair çıkarımlar yapmaktır. Bu bağlamda, klasik test kuramının ana kavramlarından "güvenirlik" yerine, genellenebilirlik kuramında daha kapsamlı ve esnek bir yaklaşım olarak "genellenebilirlik" tercih edilir (Güler, 2009).

Psikolojik ve eğitimsel ölçme alanında en yaygın kullanılan kuramlarından olan klasik test kuramı (KTK) yapılan ölçme işlemini değerlendirebilmek ve yorumlayabilmek için kullanılan bir kuramdır (Baykul, 2021). Güvenirlik ise, bir ölçme aracının arka arkaya yapılan ölçmelerde benzer sonuçlar vermesidir diğer bir ifadeyle ölçüm sonuçlarının tesadüfi hatalardan arınık olmasıdır (Tekin, 1994). Klasik test kuramı kapsamında, güvenirlik hesaplamaları yalnızca belirli bir hata kaynağına odaklanır ve güvenirliğin tanımına bağlı olarak farklılık gösterir.

Örneğin, test-tekrar test yöntemi için "zaman aralığı" hatası, paralel formlar yöntemi için "formların eşdeğerliği" hatası hesaplanır (Cohen ve Swerdlik, 2018). Bu nedenle, aynı ölçüm için klasik test kuramında farklı hata kaynaklarına dayalı hesaplanan güvenirlik katsayıları birbirinden farklı sonuçlar verebilir. Diğer taraftan genellenebilirlik kuramı, ölçüm sürecinde yer alan puanlayıcı, zaman, test formu, madde ve görev gibi tüm olası değişkenlik kaynaklarından meydana gelen hataları birlikte ve eşzamanlı olarak değerlendirme fırsatı sunmaktadır (Atılgan, 2019). Genellenebilirlik kuramının önemli avantajlarından biri de farklı hata kaynaklarının etkileşim etkilerini değerlendirebilme kabiliyetidir. Genellenebilirlik kuramı test puanları üzerindeki çoklu hata varyansı kaynaklarının birleşik, ortak etkilerini aynı anda değerlendirmek için varyans analizi (ANOVA) yöntemlerini kullanır. Bu durum, araştırmacıların puanlara etki edebilecek hata varyansı bileşenlerini tanımlayabilmeleri için daha kapsamlı bir yöntem sunar (Urbina, 2004).

Genellenebilirlik kuramında dikkate alınan olası hata kaynaklarının her biri “yüzey” (facet) olarak kabul edilir. Örneğin; maddeler (görevler), puanlayıcılar, form ve zaman vb. kaynaklar farklı yüzeyleri oluştururlar (Crocker ve Algina, 2006). Bireyler bir yüzey olarak kabul edilmemektedir. Çünkü ölçmenin temel amacı bireylere ait özelliklerin ortaya çıkarılmasıdır. Bu nedenle bireyler yüzey olarak kabul görmemektedir (Atılgan, 2019). Yüzeyler içerisinde yer alan birden fazla düzey ise “koşul” (condition) olarak adlandırılır. Puanlayıcı yüzeyi düşünüldüğünde araştırma üç farklı puanlayıcı barındırıyorsa bunların her biri bir “koşul” olarak değerlendirilir (Güler, 2009). Örneğin; 10 maddelik bir mülakat sınavı ve bu sınavı değerlendiren 4 puanlayıcı olduğunu varsayalım. Bu durumda sınavda yer alan madde ve puanlayıcı yüzey olarak kabul edilir, sınavı oluşturan her bir madde ve her bir puanlayıcı ise yüzeylere ait koşulları oluşturur. Bu durumda, bu yüzeylerde belirlenen koşullardan yapılabilecek tüm gözlemler kabul edilebilir gözlemlerin evrenini temsil eder. Diğer bir ifadeyle bu evren, birey, madde ve puanlayıcı yüzeylerinden elde edilen tüm ölçümleri ve bunlara ait varyansları kapsar. Genellenebilirlik evreni ise, genellenmek istenilen yüzeylere ait koşulların (evrendeki tüm madde, puanlayıcı vb.) hepsini kapsayan bir evrendir (Price, 2021; Güler, 2009). Bir bireyin teorik olarak sonsuz sayıda yapılan ölçme işlemine ait sonuçları üzerinden ortalama puanı ise genellenebilirlik kuramına göre o bireyin evren puanıdır. Bu puan KTK’ da yer alan “gerçek puan” kavramına eş değerdir (Price, 2021).

Ölçme işleminden kaynaklanan tüm hataları dikkate alarak, mevcut tüm hata kaynaklarını tespit etmek, belirgin etkilerini ortaya çıkarmak ve uygun kabul edilebilir gözlemlerin evrenini betimlemek için G çalışması yapılır. Bu doğrultuda amaç, ölçme sonuçlarındaki varyans

kaynaklarını belirleyerek ölçme işlemi hakkında bazı kararlar elde etmek, sonraki ölçmelerde bu kararlar ışığında gelebilecek hataları belirlemek ve bunların azaltılmasını sağlamaktır (Atılgan, 2019). G-çalışmasının diğer amacı ise K-çalışmasının yeterli genelleme kapasitesine sahip olmasına yardımcı olmaktır. K çalışmasının temel amacı ise, ölçmede ortaya çıkan hataların en aza indirgenmesi için en uygun ölçme desenin ortaya konmasını sağlamaktır (Price, 2021). Bir K-çalışmasında bir yüzey sabit ya da rastgele olarak ele alınabilir. Bir yüzey sabitse, araştırmacı yalnızca K-çalışmasında ortaya çıkan koşullara genelleme yapmayı amaçlar. Bir yüzey rastgele ise, araştırmacı K-çalışmasındaki koşulları daha fazla sayıda koşuldan bir örneklem olarak kabul eder ve tüm bu son koşullara genelleme yapmayı amaçlar (Crocker ve Algina, 2006). Eğer bir sınavda yer alan maddeler yine benzer bir alan sınavında yer alabilecek nitelikte ve o maddelere eş değer olabilirse bu maddeler rastgele olarak adlandırılır ve bu durumda bu maddeler evrende yer alan tüm maddelere benzer varyans kaynaklarına sahip olabileceği için evrene genelleme yapılabilmesine olanak sağlar. Diğer taraftan araştırmada sadece belli durumlara yönelik genelleme amacı varsa bu durumda varyans kaynağı sabit olarak kabul edilir (Güler, 2009). Değişkenlik (varyans) kaynakları, ölçüm hatalarının değerlendirilmesi ve bu hataların ölçüm sürecine etkisinin analiz edilmesinde kritik bir rol oynar. G kuramında, bu değişkenlik kaynakları farklı desenler halinde modellenmektedirler (Brennan, 2001).

Genellenebilirlik kuramıyla, elde edilen ölçme sonuçlarının değerlendirilmesi sürecinde çaprazlanmış (crossed) veya yuvalanmış (nested) bir yapı-desen oluşturulabilir (Atılgan, 2019). Çapraz desenlerde, ölçmeye tabi tutulan bireyler mevcut tüm yüzeylerin her koşulunda puanlanmaktadır. Örneğin, açık uçlu bir sınavı alan her birey (b) ölçme aracında yer alan bütün (m) maddeleri cevaplamış ve üç farklı (p) puanlayıcı her bireyin tüm yanıtlarına yönelik puanlama yapmışsa, bireyler, maddeler ve puanlayıcılar çaprazlanmış bir deseni oluşturacaklardır (Güler, 2009). Bu durumda yüzeyler arasına “x” işareti konularak bir gösterim yapılır: “b x m x p”. Yuvalanmış desen ise, bir yüzeye ait koşullar diğer bir yüzeye ait tüm koşullar tarafından gözlenirse ve bir yüzeyin sadece bazı koşulları diğer bir yüzeyin sadece birkaç koşulunca gözlenirse ortaya çıkar. Burada bireyler yüzeylerin sadece bir koşulu için puanlanmışken diğer koşullarda puanlanması bulunmamaktadır. Örneğin, bir açık uçlu sınavı alan her (b) birey ölçme aracında yer alan bütün (m) maddeleri cevaplamış ancak her maddeye verilen cevaplara farklı (p) puanlayıcılar puanlama yapmışsa burada kısmi yuvalanmış desen söz konusudur. Bu durumda desen şu şekilde gösterilir: “b x (p : m)” (Atılgan, 2019).

Proje yarışmalarında kullanılan değerlendirme süreçleri, projelerin geçerli ve güvenilir bir şekilde sıralanabilmesi için önemlidir. Benzer konu alanıyla ilgili mevcut alanyazın incelendiğinde, öğrenci performanslarını değerlendirmeye yönelik puanlayıcı güvenirliğini genellenebilirlik kuramına dayalı olarak inceleyen bazı çalışmalar bulunmaktadır (Büyükkıdık ve Anıl, 2015; Lafave ve Butterwick, 2014; Aktaş ve Alıcı, 2017; Menéndez-Varela vd., 2018; Gülle vd., 2018; Stuhlmann vd., 1999; Taştan ve Uzun, 2021; Özbaşı ve Arcagök, 2021; Tindal vd., 2010; VanLeeuwen, 1997). Yapılan bu araştırmaların bazılarında, çapraz ve yuvalanmış farklı desenler kullanılarak puanlayıcı güvenirliğini artırmaya yönelik desen karşılaştırmaları amaçlanmış, bazılarında farklı puanlama anahtarları kullanılarak uygun puanlama ve puanlayıcı sayısına ulaşılmaya çalışılmıştır. Bazı araştırmalar ise öğrenci performanslarını ölçmeye yönelik güvenilir puanlama anahtarı geliştirmek amacıyla G kuramı kullanımı üzerine yapılmıştır. Ancak özelinde proje yarışmalarının değerlendirme süreçlerini dikkate alan bir çalışma mevcut alanyazında görülmemiştir.

Ulusal ve uluslararası düzeyde bilişim teknolojileri temelli yarışmaların değerlendirme süreçleri incelendiğinde, puanlayıcıların genellikle dereceli puanlama anahtarları kullandığı görülmektedir. Ancak, bu anahtarların içerdiği farklı boyutlar (örneğin, teknik konular ve proje yazım kuralları) açısından güvenilirliği ve puanlayıcılar arasındaki değerlendirme farklılıklarının etkisi belirsizliğini korumaktadır. Projeler, genellikle konu alanıyla ilgili ikiden fazla değerlendirici tarafından puanlanmakta ve bu puanların ortalaması alınarak genel bir değerlendirme notu oluşturulmaktadır. Bununla birlikte, yalnızca proje konusu uzmanı olan değerlendiriciler puanlama yaparken projenin fikri ve uygulaması güçlü olduğu durumlarda proje yazım kurallarını göz ardı edebilmektedir. Bu durum, teknoloji temelli proje yarışmalarında güvenilir ve genellenebilir değerlendirme süreçlerinin oluşturulması için farklı desenlerin incelendiği bir çalışmaya duyulan ihtiyacı ortaya koymaktadır. Bu bağlamda, projelerin geçerli ve güvenilir bir şekilde değerlendirilmesini sağlamak amacıyla, proje yazım kurallarını proje yazım uzmanlarının, teknik konularla ilgili kriterleri ise teknik konu uzmanların değerlendirmesi daha güvenilir sonuçlara ulaşılmasını mümkün kılabilir.

Genellenebilirlik kuramı, bu tür çok boyutlu ve çok puanlayıcılı değerlendirme süreçlerindeki varyans kaynaklarını ve güvenilirlik düzeylerini analiz etmek için güçlü bir yöntem sunmaktadır. Bu nedenle bu kurama dayalı yaklaşımların, bu alandaki boşluğu doldurabileceği ve değerlendirme süreçlerini daha şeffaf bir yapıya kavuşturabileceği düşünülebilir. Bu bağlamda yapılan bu araştırmanın amacı, proje değerlendirme süreçlerinde genellenebilirlik kuramına dayalı farklı desenler oluşturarak bu süreçler

üzerinde farklı puanlayıcıların etkisini G kuramı bağlamında ortaya koymaktır. Sonuç olarak bu metodolojik araştırmada oluşturulan desenler ile bilişim teknolojilerini temele alan proje yarışmalarının değerlendirme süreçlerinin iyileştirilmesine katkıda bulunulmak istenmektedir. Bu amaç doğrultusunda aşağıdaki araştırma sorularına yanıt aranmaktadır:

1- Proje (b), puanlayıcı (p) ve madde (m) yüzeylerinin çapraz desen (b x m x p) oluşturduğu değerlendirme sürecine yönelik elde edilen G ve Phi katsayıları ile öne çıkan varyans kaynakları nelerdir?

2- Puanlayıcıların (p) madde (m) yüzeyinde yuvalandığı ve projelerin (b) bu iki yüzey ile çapraz desen (b x (p : m)) oluşturduğu değerlendirme sürecine yönelik elde edilen G ve Phi katsayıları ile öne çıkan varyans kaynakları nelerdir?

3- Proje (b), puanlayıcı (p) ve madde (m) yüzeylerinin çapraz desen (b x m x p) oluşturduğu değerlendirme süreci ile puanlayıcıların (p) madde (m) yüzeyinde yuvalandığı ve projelerin (b) bu iki yüzey ile çapraz desen (b x (p : m)) oluşturduğu değerlendirme süreci sonuçlarına göre elde edilen G ve Phi katsayıları birbirlerinden farklılaşmakta mıdır?

Yöntem

Araştırmanın Türü

Bu araştırma, genellenebilirlik kuramına dayalı farklı desenlerin değerlendirme sürecinin güvenirliği üzerindeki etkisini inceleyen metodolojik bir çalışmadır. Araştırma, gerçek proje yarışmalarının değerlendirme yapısını yansıtacak şekilde tasarlanmış simüle veri seti kullanılarak yürütülmüştür. Simülasyon yaklaşımının tercih edilmesinin başlıca nedenleri şunlardır: (1) Kontrollü karşılaştırma: Farklı desenlerin etkilerini diğer değişkenlerin etkisinden arındırarak incelemek, (2) Sistematik manipülasyon: Puanlayıcı ve madde sayılarını sistematik olarak değiştirerek optimal koşulları belirlemek, (3) Genellenebilirlik kuramının farklı desenlerinin teorik özelliklerini net bir şekilde ortaya koymak, (4) Gerçek yarışma organizatörlerine değerlendirme sistemi tasarımları için kanıta dayalı öneriler sunmak. Simüle edilen veri seti, TÜBİTAK ve benzeri ulusal düzeydeki proje yarışmalarının gerçek değerlendirme süreçlerini (puanlayıcı sayısı, madde yapısı, puanlama ölçeği vb.) yansıtacak şekilde farklı puanlama desenlerine uygun olarak tasarlanmıştır. Bu tarz yaklaşımlar, metodolojik karşılaştırmalarda ve ölçme kuramlarının uygulanmasında yaygın olarak kullanılmaktadır.

Veri Üretme Süreci

Bu çalışmada kullanılan simüle veri seti, R dili kullanılarak R Studio 2023.09.0 paket programı ile oluşturulmuştur. Her proje için normal dağılıma dayalı değerlendirme puanlarına ait kontrollü bir veri seti oluşturulmuştur. Bu şekilde elde edilen veriye gerçekçi bir varyasyon kazandırılmaya çalışılmıştır. Puanlayıcıların her maddeye verdiği kabul edilen cevaplara ait oluşan simüle puanlar, 1–5 aralığında en yakın tam sayıya yuvarlanmıştır. Bu süreç, proje(b), madde(m) ve puanlayıcı(p) farklılıkları ile bunların karşılıklı etkileşimlerinden kaynaklanan varyans bileşenlerini yansıtacak biçimde tasarlanmıştır.

Simülasyonun, gerçek proje değerlendirme süreçlerini ne ölçüde yansıttığını belirlemek amacıyla çeşitli doğrulama adımları da uygulanmıştır. Elde edilen değerlerin benzer araştırmalarda rapor edilen bulgularla karşılaştırılarak sonuçların tutarlılığı değerlendirilmiştir (Özbaşı ve Arcagök, 2021; Taştan ve Uzun, 2021). Ayrıca, elde edilen simüle veri, normal dağılım varsayımı ve varyans çeşitliliği açısından kontrol edilmiştir. Sonuç olarak simülasyonun gerçek değerlendirme koşullarını istatistiksel açıdan güvenilir biçimde yansıttığı düşünülmektedir.

Proje Değerlendirme Aracı

Simülasyonda, TÜBİTAK 2204-B Ortaokul Öğrencileri Araştırma Projeleri Yarışması Teknolojik Tasarım Alanı Değerlendirme Formu'ndaki madde yapısı kullanılmıştır (Tablo 1). Bu form 10 farklı başlık altında 20 kriter içermektedir. Her madde 1(Çok düşük) - 5(Çok yüksek) arası puanlanmaktadır.

Puanlama Desenleri

Projelerin değerlendirilmesi için genellenebilirlik kuramına dayalı iki farklı puanlama deseni tasarlanmıştır:

İlk desende, bilişim teknolojileri alanında uzman ilk üç puanlayıcı 10 adet projenin tümünü dereceli puanlama anahtarı ile puanladığı varsayılmıştır. Burada oluşan çaprazlanmış desen şu şekildedir: “b x m x p”.

Tablo 1.

Veri toplama aracı olarak kullanılan değerlendirme formu

2204-A Lise Öğr. Araş. Proj. Yarışması ve 2204-B Ortaokul Öğr. Araş. Proj. Yarışması

Teknolojik Tasarım Alanı Ön Değerlendirme Formu

1	2	3	4	5
Çok Düşük	Düşük	Yeterli	Yüksek	Çok Yüksek

Kriter ve Göstergeler	1	2	3	4	5
Başarılabilirlik (15p)					
1. Kaynakların ekonomik kullanım düzeyi					
2. Problemi çözümlemek için alt amaçların uygun şekilde formüle edilme düzeyi					
3. Bulguların prototipin başarısını kanıtlama düzeyi					
Yaratıcılık (15p)					
4. Teknolojik Tasarım alanına katkı sağlama potansiyeli					
5. Teknolojik Tasarım alanına farklı bir perspektif getirme potansiyeli					
6. Alan uzmanlarının ilgisini çekme düzeyi					
Özgünlük (15p)					
7. Daha önce yapılmış çalışmalardan farklılık düzeyi					
8. Teknolojik Tasarım alanındaki bilgi birikimini ileri taşıma potansiyeli					
9. Teknolojik Tasarım alanında yeni gelişmelere yol açma potansiyeli					
Etik İlkelere Uygunluk (10p)					
10. Alanyazın taramasının yeterlik düzeyi					
11. Atıfların metin içinde ve kaynakçada rehbere uygunluk düzeyi					
Sonuç ve Öneri (10p)					
12. Raporlanan sonuçların tekrarlanabilirlik düzeyi					
13. Sonuç ve önerilerin açıklık ve anlaşılırlık düzeyi					
Hedef Kitle (5p)					
14. Hedef kullanıcıların temel özelliklerinin tanımlanma düzeyi					
Müdahale / Ürün (5p)					
15. Geliştirilen ürünün farklı koşullarda çalışabilme potansiyeli					
Karşılaştırma (10p)					
16. Projenin güçlü yönlerinin ikna edicilik düzeyi					
17. Projenin geliştirilebilir yönlerinin tespit edilme düzeyi					
Çıktı (10p)					
18. Uygulama alanına metodolojik/kavramsal/kuramsal olarak katkıda bulunma potansiyeli					
19. Öngörülen sosyal/ekonomik etkilerin ulaşılabilirlik düzeyi					
Zaman Yönetimi (5p)					
20. Proje metninin rehbere uygunluğu					

İkinci desende ise tüm projeler için şu şekilde bir değerlendirme yapıldığı varsayılmıştır: Öncelikle değerlendirme formunda (Tablo 1) yer alan maddeler, uzman görüşleri doğrultusunda hangi maddeleri proje yazım uzmanı hangi maddeleri teknik konu uzmanı puanlayıcıların yanıtlamaları gerektiği konusunda iki gruba ayrılmıştır. Buna göre her projede,

proje yazımı konusu uzmanı olan üç puanlayıcı kendi uzmanlık alanları ile ilgili 10, 11, 12, 13, 14, 18 ve 20. maddeler olmak üzere toplam 7 maddeyi puanlamış, ilk aşamada değerlendirme yapmayan üç bilişim teknolojileri alanı uzmanı puanlayıcı ise kendi uzmanlık alanları ile ilgili 1, 2, 3, 4, 5, 6, 7, 8, 9, 15, 16, 17, 19. maddeler olmak üzere her proje için toplam 13 maddeyi puanlamıştır. Burada ise puanlayıcıların maddelerde yuvalandığı ve bireylerin maddeler ve puanlayıcılarla çaprazlandığı bir desen "b x (p : m)" tasarlanmıştır. Bu süreçte iki farklı türdeki tüm puanlayıcılar için oluşturulan puanlar her projenin toplam puanına etki etmiştir.

Verilerin Çözümlenmesi

Simülasyon ile elde edilen verilerin çözümlenmesi için çapraz ve kısmi yuvalanmış desene ait değişkenlik kaynakları için kareler ortalaması, varyans bileşenleri değerleri ve yüzdelik oranları ANOVA testi ile hesaplanmıştır. Devamında iki farklı desene ait G ve Phi katsayıları ayrı ayrı hesaplanarak incelenmiştir. Karar çalışmaları için farklı madde ve puanlayıcı sayılarındaki katsayılar da hesaplanarak en uygun desene hangisi olabileceği incelenmiştir. İstatistiksel hesaplamaların yapılabilmesi için IBM SPSS 27.0 paket programı ve RStudio 2023.09.0 programı kullanılmış, elde edilen veriler tablolara aktarılmıştır.

Etik Kurul İzin Belgesi

Bu çalışma metodolojik bir simülasyon çalışması olup, gerçek katılımcı verisi içermemektedir. Bu nedenle etik kurul onayı gerektirmemektedir.

Bulgular

Bu bölümde araştırma için simülasyon ile oluşturulan iki farklı desene ait analizler sonucunda elde edilen bulgular verilmiştir.

Birinci Alt Probleme ait Bulgular

G Çalışması Bulguları

Bilişim teknolojileri alanında uzman ilk üç puanlayıcı 10 adet projenin tümünü dereceli puanlama anahtarı ile puanlamıştır. Burada oluşan çaprazlanmış desen şu şekildedir: 10 Proje (b), 3 puanlayıcı (p) ve 20 madde (m) kullanılarak oluşturulan iki yüzeyli çaprazlanmış desen (b x m x p). Bu araştırma desenine yönelik genellenebilirlik çalışmaları yapılmış ve bu analiz sonucunda elde edilen G çalışması sonuçları tablo 2'de verilmiştir.

Tablo 2.
İki yüzeyli çapraz desene (bxmxp) ait G çalışması sonuçları

Varyans Kaynağı	Kareler Toplamı	sd	Kareler Ortalaması	Varyans	%
Proje (b)	57.35	9	6.37	0.087	8.1
Madde (m)	168.56	19	8.87	0.274	25.7
Puanlayıcı (p)	110.26	2	55.13	0.270	25.3
Proje * Madde (b*m)	87,552	171	0,51	0,059	5.6
Proje * Puanlayıcı (b*p)	17,437	18	0,97	0,032	3.0
Madde*Puanlayıcı (m*p)	17,470	38	0,46	0,013	1.2
bmp, e	114,163	342	0,33	0,334	31.2
TOPLAM	5171,000	600		1,069	100

Tablo 2 incelendiğinde projelerin (b) ana etkisinin toplam varyansın %8.1'i olduğu görülmektedir. Bu durum, projelerin arasındaki farklılıkların yapılan çalışma ile az bir miktarda ortaya konabildiğini göstermektedir. Diğer bir ifade ile yapılan puanlamaların projeleri birbirinden ayırt etmede çok da başarılı olmadığı görülmektedir. Buna göre elde edilen puanların evren (gerçek) puanları temsil etmede yeterli olmadığı anlaşılmaktadır.

Maddelerin (m) ana etkisine bakıldığında elde edilen madde varyans değerinin ikinci en büyük varyans (0.274) olduğu ve toplam varyansın %25.7'sini açıkladığı görülmektedir. Buna göre yapılan değerlendirme işlemi üstünde maddelerin büyük bir etkisinin olduğu ve maddelerin güçlük değerleri arasında bir farklılık olduğu anlaşılmaktadır.

Değerlendirmeyi yapan puanlayıcıların ana etkisi incelendiğinde varyans değerinin en büyük üçüncü paya (%25.3) sahip olduğu ve toplam varyans içinde büyük bir etkisinin olduğu görülmektedir. Bu durum puanlayıcıların yaptıkları puanlamalarda katılık ya da cömertlik açısından birbirlerinden farklı davrandıkları ve puanlamalar arasında projeden projeye farklılık olduğunu göstermektedir.

Proje ve madde (b*m) ortak etkisinin varyans değeri (0.059) toplam varyansın % 5.6' sını açıklamaktadır. Bu durum projelerin farklı maddelerde az da olsa farklı puanlar aldıklarını ve projelerin farklı maddelerde farklı sıralamalara sahip olduğunu göstermektedir.

Proje ve puanlayıcı (b*p) ortak etkisine bakıldığında varyans değerinin düşük (0.032) olduğu ve toplam varyans içinde yine düşük bir oran (%3) oluşturduğu görülmektedir. Buna göre puanlayıcıların projeleri puanlarken düşük seviyede farklılık gösterdiği ve genel olarak birbirlerine yakın puanlamalar yaptığı söylenebilir.

Madde ve puanlayıcı (m*p) ortak etkisinin yine düşük olduğu (0.013) ve toplam varyansın %1.2' sini açıkladığı görülmektedir. Bu durum, puanlayıcıların her proje için maddelere verdikleri puanlamalarda birbirlerine yakın davranışlar sergiledikleri ve madde puanlamaları arasında bir farklılığın olmadığı anlamına gelmektedir.

Tüm varyans kaynaklarının ortak etkisi olan (bmp, e) artık varyans %31.2 ile en büyük varyans kaynağıdır. Bu varyans değeri (0.334) proje, madde ve puanlayıcı arasında oluşturulan desen ile açıklanamayan hata varyansını göstermektedir. Artık varyansın bu denli yüksek olması, yapılan çalışmada ölçülememiş tesadüfi hata kaynaklarının yüksek düzeyde var olduğunu işaret etmektedir.

K Çalışması Bulguları

Çalışmada elde edilen 10 proje, 3 puanlayıcı ve 20 maddenin çapraz desenine yönelik G ve Phi katsayıları tablo 3'te verilmiştir.

Tablo 3.
Çapraz Desene (bxmxp) ait Karar çalışması sonuçları

Puanlayıcı Sayısı (n_p)	Madde Sayısı (n_m)	G Katsayısı	Phi Katsayısı
2	10	0.90	0.44
3	10	0.94	0.44
4	10	0.94	0.44
2	20	0.94	0.44
***3**	***20**	***0.96**	***0.45**
4	20	0.96	0.45
2	30	0.95	0.45
3	30	0.97	0.45
4	30	0.97	0.45

Buna göre 3 puanlayıcı ve 20 madde için G katsayısının 0.96 olarak hesaplandığı görülmektedir. Bu değer bağıl hata varyansı olarak da bilinmekte olup .70 ve üzeri olduğu durumlarda çalışmanın güvenilir olduğu kabul edilmektedir. Burada hesaplanan değer .70' den büyük olduğundan ölçmenin bağıl kararlar için güvenirliğinin çok yüksek olduğunu göstermektedir.

Tablo 3' te görülen 3 puanlayıcı (n_p=3) ve 20 madde (n_m=20) için phi (Φ) katsayısı, 0.45 olarak hesaplanmıştır. Bu değer mutlak hata varyansı olarak da bilinmekte olup G katsayısı gibi bu değer de .70 ve üzeri olduğu durumlarda çalışmanın güvenirliği yeterli düzeyde kabul edilmektedir (Shavelson ve Webb, 1991). Ancak buradaki phi (Φ) katsayısı değerinin .70' den düşük olduğu görülmekte ve bu nedenle çalışmanın mutlak değerlendirme için güvenirliğinin

düşük düzeyde olduğu anlaşılmaktadır.

Tablo 3 ayrıntılı olarak incelendiğinde, proje, madde ve puanlayıcıların çaprazlanmış deseninde puanlayıcı ve madde sayısı artırıldıkça güvenirlik değerlerinin de arttığı görülmektedir. Tabloda yer alan en yüksek güvenirlik değerleri 4 puanlayıcı ve 30 madde kullanıldığında elde edildiği görünse de mevcut araştırmada kullanılmış çapraz desene ait değerlerin bunlara yakın olduğu ve karar çalışması için yeterli olduğu görülmektedir.

İkinci Alt Probleme ait Bulgular

G Çalışması Bulguları

Çalışmanın ikinci simülasyon deseninde, proje yazımı uzmanı olan üç puanlayıcı kendi uzmanlık alanları ile ilgili her projedeki 10, 11, 12, 13, 14, 18 ve 20. maddeler olmak üzere toplam 7 maddeyi puanlamış, ilk aşamada değerlendirme yapmayan diğer üç bilişim teknolojileri alanı uzmanı puanlayıcı ise kendi uzmanlık alanları ile ilgili her projedeki 1, 2, 3, 4, 5, 6, 7, 8, 9, 15, 16, 17, 19. maddeler olmak üzere toplam 13 maddeyi puanlamıştır. 6 puanlayıcının (p) 20 madde (m) yüzeyinde yuvalandığı ve 10 projenin (b) bu iki yüzey ile çapraz desen (b x (p : m)) oluşturduğu değerlendirme sürecine yönelik elde edilen G çalışması sonuçları tablo 4' te verilmiştir.

Tablo 4.
İki yüzeyli kısmi yuvalanmış desene (bx(p:m)) ait G çalışması sonuçları

Varyans Kaynağı	Kareler Toplamı	sd	Kareler Ortalaması	Varyans	%
Proje (b)	57.35	9	6.37	0.071	6.9
Puanlayıcı (p)	164.27	5	32.85	0.322	31.4
Proje * Puanlayıcı(b*p)	45.99	45	1.02	0.069	6.7
Puanlayıcı : Madde(p:m)	132.03	54	2.44	0.209	20.4
b*p:m, e	173.16	486	0.36	0.354	34.6
TOPLAM	572,80	599	-	92.096	100

Tablo 4 incelendiğinde projelerin (b) ana etkisinin toplam varyans içinde %6.9'luk küçük bir kısmı açıkladığı görülmektedir. Bu bulguya göre projelerin arasındaki farklılıkların yapılan değerlendirme ile yeterince ortaya çıkarılamadığı görülmektedir. Diğer bir ifade ile yapılan puanlamaların projeleri birbirinden ayırt etmede çok da başarılı olmadığı görülmektedir. Buna göre elde edilen puanlar, evren (gerçek) puanlarını temsil etmede yeterli değildir.

Puanlayıcıların oluşturduğu ana etki değerine bakıldığında bu değerin (0.322) en yüksek

ikinci değer olduğu ve toplam varyansın %31.4'ünü açıkladığı görülmektedir. Buna göre puanlayıcılar araştırma deseni içerisinde büyük bir etki oluşturmuştur. Bunun nedeni olarak puanlayıcıların birbirlerinden farklı puanlama davranışları sergiledikleri gösterilebilir. Sonuç olarak bu desenle yapılan çalışmada puanlayıcıların tutarlı ve birbirlerine yakın davranmadığı bir durum söz konusudur.

Proje ve puanlayıcıların (b * p) ortak etkisine ait varyans değerinin (0.069) küçük bir miktar olduğu ve toplam varyans içerisinde (%6.7) küçük bir bölümü açıkladığı görülmektedir. Bu bulgu, puanlayıcıların projeleri puanlarken projeden projeye düşük seviyede farklı davrandıkları ve projelere verdikleri puanlar arasında bir miktar farklılık olduğu sonucunu vermektedir.

Madde ve puanlayıcıların (p : m) ortak etkisine bakıldığında elde edilen varyans değerinin (0.209) en yüksek üçüncü varyans değeri olduğu ve toplam varyans içinde yine yüksek bir bölümü (%20.4) açıkladığı görülmektedir. Bu bulguya göre, puanlayıcıların puanlama davranışlarının bir maddeden diğerine değiştiği söylenebilir.

Tüm değişkenlik kaynaklarından ortaya çıkan ortak hata varyansı (bxp:m,e) incelendiğinde bu değerin en yüksek varyans bileşeni (0.354) olduğu ve toplam varyansın %34.6'sını açıkladığı görülmektedir. Bu bulguya göre araştırmanın bu deseni ile yapılan ölçme işlemine proje, madde ve puanlayıcıdan kaynaklanmayan tesadüfi hataların yüksek düzeyde karıştığı anlaşılmaktadır.

K Çalışması Bulguları

Çalışmada elde edilen 10 proje, 6 puanlayıcı ve 20 maddenin kısmi yuvalanmış desenine ait G ve Phi katsayıları tablo 5' te verilmiştir.

Tablo 5.
İki Yüzeyli Kısmi Yuvalanmış Desene (bx(p:m)) ait karar çalışması sonuçları

Puanlayıcı Sayısı (n_p)	Madde Sayısı (n_m)	G Katsayısı	Phi Katsayısı
4	10	0.82	0.21
6	10	0.85	0.16
8	10	0.86	0.13
4	20	0.90	0.33
***6**	***20**	***0.92**	***0.27**
8	20	0.93	0.23

Puanlayıcı Sayısı (n_p)	Madde Sayısı (n_m)	G Katsayısı	Phi Katsayısı
4	30	0.93	0.39
6	30	0.94	0.34
8	30	0.95	0.30

Tablo 5'e göre çalışmada yer alan 6 puanlayıcı ve 20 madde içeren desen için G katsayısının 0.92 olarak hesaplandığı görülmektedir. Bu değer bağıl hata varyansı olarak da bilinmekte olup .70 ve üzeri olduğu durumlarda çalışmanın güvenilir olduğu kabul edilmektedir. Burada hesaplanan değer eşik değerden (>.70) olduğundan ölçmenin bağıl kararlar için güvenirliğinin çok yüksek olduğunu göstermektedir.

Çalışmada kullanılan ve tablo 5' te görülen 6 puanlayıcının (n_p=6) 20 madde (n_m=20) içinde yuvalandığı desene ait phi (Φ) katsayısı, 0.27 olarak hesaplanmıştır. Bu değer mutlak hata varyansı olarak da bilinmekte olup G katsayısı gibi bu değer de .70 ve üzeri olduğu durumlarda çalışmanın güvenirliğinin yeterli düzeyde olduğu kabul edilmektedir. Ancak buradaki phi (Φ) katsayısı değerinin .70' den düşük olduğu görülmekte ve bu nedenle çalışmanın mutlak değerlendirme için güvenirliğinin yeterli düzeyde olmadığı görülmektedir.

Tablo 5 ayrıntılı olarak incelendiğinde, puanlayıcıların maddelerde yuvalandığı bu desende puanlayıcı ve madde sayısı artırıldıkça güvenirlik değerlerinin de genel olarak arttığı görülmektedir. Tabloda yer alan en yüksek güvenirlik değerleri 8 puanlayıcı ve 30 madde (G= 0.95, phi=0.30) kullanıldığı zaman elde edildiği görünse de mevcut araştırmada kullanılmış kısmi yuvalanmış desene ait değerlerin (G= 0.92, phi=0.27) bunlara yakın olduğu ve karar çalışması için yeterli olduğu görülmektedir.

Üçüncü Alt Probleme ait Bulgular

Proje (b), puanlayıcı (p) ve madde (m) yüzeylerinin çapraz desen (b x m x p) oluşturduğu ölçme işlemi süreci ile puanlayıcıların (p) madde (m) yüzeyinde yuvalandığı ve projelerin (b) bu iki yüzey ile çapraz desen (b x (p : m)) oluşturduğu ölçme işlemi süreci sonuçlarına göre elde edilen G ve Phi katsayıları ayrı ayrı incelenmiştir.

Elde edilen bulgulara göre, tüm yüzeylerin çapraz desen oluşturduğu ilk çalışmaya ait G katsayısı 0.96 iken kısmi yuvalanmış desene ait G katsayısının 0.92 olduğu görülmektedir. Bu bulgular, iki desenle de yapılan ölçme işleminin bağıl değerlendirme açısından güvenilir seviyede olduğunu (>.70) göstermekle birlikte çapraz desene ait güvenirlik değerinin bir

miktar daha yüksek olduğunu göstermektedir.

Genellenebilirlik analizi sonucunda ulaşılan tüm yüzeylerin çapraz desen oluşturduğu ilk çalışmaya ait phi katsayısı 0.45 iken kısmi yuvalanmış desene ait phi katsayısı ise 0.27'dir. Bu bulgular, her iki desenin de mutlak değerlendirme açısından çok da güvenilir düzeyde olmadığını ancak çapraz desene ait phi katsayısı değerinin kısmi yuvalanmış desene oranla daha güvenilir olduğunu ortaya koymaktadır.

Tartışma ve Sonuç

Bu metodolojik araştırmanın amacı, proje değerlendirme süreçlerinde genellenebilirlik kuramına dayalı farklı desenler oluşturarak bu süreçler üzerinde farklı puanlayıcıların etkisini G kuramı bağlamında ortaya koymaktır. Araştırmada uygulanan ilk simülasyon desende bilişim teknolojileri alanında uzman olduğu varsayılan ilk üç puanlayıcı 10 adet projenin tümünü dereceli puanlama anahtarı ile puanlamıştır. Bu ölçme işleminde yer alan 10 proje (b), 3 puanlayıcı (p) ve 20 madde (m) kullanılarak oluşturulan iki yüzeyli çaprazlanmış desene (bxmxp) ait genellenebilirlik çalışmaları yapılmıştır. İkinci simülasyon desende üç bilişim teknolojileri alan uzmanı ve üç proje yazım uzmanı olduğu varsayılan toplam 6 puanlayıcının farklı sayılardaki maddeler (toplam 20 madde) üzerinde yuvalandığı ve 10 projenin bu iki yüzey ile çapraz desen (bx(p:m)) oluşturduğu değerlendirme sürecine yönelik genellenebilirlik çalışmaları yapılmıştır. İki farklı desenden elde edilen sonuçlar şu şekilde özetlenmiştir:

Her iki desende de projelerin ana etkisinin düşük çıkması (çaprazlanmış desen %8.1, yuvalanmış desen %6.9), tasarlanan ölçme işlemlerinin projeleri yeterince ayırt edemediğini ve yapılan ölçme işlemlerinin projelerdeki nitelik farklarını yeterince yansıtamadığını göstermektedir. Bu bulgu, Özbaşı ve Arcagök (2021)'ün öğretmen adaylarının hazırlamış oldukları projelerin bağımsız puanlayıcılar ile puanlanarak elde edilen ölçümlerin genellenebilirlik kuramı ile incelenmesine ilişkin çalışmasındaki bulgularla (%0,4) örtüşmektedir. Benzer şekilde, Sun ve diğerlerinin (1997) çalışmasında da projelere ait varyans değerinin diğerlerine oranla daha düşük seviyede olduğu (%16.3) ve birbirlerinden tam anlamıyla ayırt edilemediği görülmüştür. Proje tabanlı ölçümlerde değişkenliğin ana kaynağının projeler, diğer bir ifadeyle ölçümün odak konusunun projeler olması öngörülmektedir (Gülle vd., 2018) ve G kuramı ile gerçekleştirilen çalışmalarda, ölçümün temel nesnesini oluşturan yüzeyin en yüksek düzeyde değişkenlik göstermesinin beklendiği belirtilmektedir (Güler, 2008). Ancak yapılan bu araştırmada yer alan ölçmenin temel nesnesi

olan projeler bu değişkenlik düzeyini yakalayamamıştır.

Çaprazlanmış desenden elde edilen bulgularda (m: %25.7) maddelerin güçlük değerleri arasında önemli farklılıklar görülmesi, Gelbal ve Çakıcı Eser (2012) tarafından test maddelerine ait varyans değerinin %16.2 olarak bulunduğu çalışmaya ait bulguyla benzerdir. Bu durum öğrencilerin proje yazım ve uygulama süreçlerinde zorlandıklarını göstermektedir. Chang ve Tseng (2011) araştırmalarında, öğrencilerin proje planlama, metodoloji geliştirme, uygulama süreci ve değerlendirme aşamalarında değişen düzeylerde zorluklarla karşılaştıklarını ortaya koymuştur. Yine Göloğlu Demir (2019)'in çalışmasında elde ettiği Tübitak araştırma projeleri yarışması ve proje hazırlama sürecine yönelik öğretmen görüşlerine göre öğrencilerin proje yazımı için gerekli yeterliğe sahip olmadıklarını ve süreç içinde zorlandıklarını belirlemiştir. Bu sonuçlar, proje hazırlama sürecinin öğrenciler için karmaşık ve zor bir süreç olduğunu desteklemektedir.

Puanlayıcı etkisinin her iki desende de yüksek çıkması, değerlendirme sürecinde puanlayıcı kaynaklı hataların önemli rol oynadığını göstermektedir. Bu sonuç, Taştan ve Uzun (2021)'un puanlayıcı eğitiminin önemine (puanlayıcı: %26,6) vurgu yaptığı çalışmasıyla örtüşmektedir. Bu bağlamda farklı çalışmalarda görüldüğü gibi (Tindal vd., 2010; Anıl ve Büyükkıdık, 2015; Doğan ve Anadol, 2016; VanLeeuwen, 1997) puanlayıcılara çalışma öncesinde nasıl puanlama yapılması gerektiği ile ilgili bir eğitim verilmesi daha iyi sonuçların elde edilebilmesini sağlayabilir. Nitekim Stuhlmann ve diğerleri (1999) araştırmalarında eğitimli puanlayıcılara ait varyans değerini %6, eğitimli olmayanlara ait varyans değerini %15 olarak tespit etmişler ve puanlayıcı eğitiminin değerlendirme tutarlılığı üzerindeki olumlu etkisini ortaya koymuşlardır. Bu sonuçlar çalışma öncesinde puanlayıcılara detaylı puanlama eğitimi verilmesinin, değerlendirme güvenirliğini artırabileceğinin göstergesidir.

Her iki desende ortaya çıkan tüm değişkenlik kaynaklarına ait ortak hata varyansı (bxp:m,e ve bmp,e) incelendiğinde bu değerlerin en yüksek varyans bileşeni (%34.6 ve %31.2) olduğu ve toplam varyansın büyük kısmını açıkladığı görülmektedir. Bu bulgular araştırmada yapılan her iki ölçme işlemine proje, madde ve puanlayıcıdan kaynaklanmayan tesadüfi hataların yüksek düzeyde karıştığını göstermektedir. Benzer şekilde VanLeeuwen (1997) en yüksek değişkenlik kaynağı olarak ortak hata varyansını %32.04, Yin ve Shavelson (2008) ise araştırmalarındaki iki farklı ölçme aracı için ortak hata varyanslarını %62.4 ve %50.3 olarak bulmuşlardır. Diğer yönden çaprazlanmış desenin yuvalanmış desene oranla daha düşük bir ortak hata varyansına sahip olduğu düşünüldüğünde puanlama anahtarındaki maddelerin

farklı uzmanlık alanlarındaki puanlayıcılar tarafından puanlanmasının daha çok hataya yol açtığı ve bu yöntem uygulandığında ölçme işlemine daha çok hata karıştığı sonucuna ulaşılabilir. Bu sonucu çaprazlanmış desenin yuvalanmış desene göre daha yüksek güvenirlik katsayılarına sahip olması da desteklemektedir.

Yapılan karar çalışması ile elde edilen sonuçlara göre proje, madde ve puanlayıcıların çaprazlanmış ve yuvalanmış deseninde puanlayıcı ve madde sayısı arttırıldıkça güvenirlik değerlerinin de arttığı görülmektedir. Bu sonuç, Lafave ve Butterwick (2014), VanLeeuwen (1997), Menendez-Varela ve Gregori-Giralt (2017)'a ait çalışmalar ile benzerlik göstermektedir. Çaprazlanmış desende kullanılan 3 puanlayıcı ve 20 madde yüksek güvenirlik değerleri verse de bu değerlere benzer 2 puanlayıcı ve 10 madde sayısının kullanılması da yüksek seviyede güvenirliği sağlamaktadır. Sonuç olarak puanlama işleminin daha az iş gücü ile, daha kısa zaman ve daha ekonomik bir şekilde yapılabilmesi için 2 puanlayıcı 10 madde sayısı da kullanılabileceği söylenebilir. Yuvalanmış desen uygulamasına bakıldığında ise 6 puanlayıcı ve 20 madde yüksek güvenirlik değerleri sağlasa da yine daha az sayıda (4) puanlayıcı ve 20 madde kullanıldığında ölçme işlemi güvenilir bir şekilde yapılabilmektedir. Sonuç olarak bu desen için de daha az iş gücü kullanılması ve ekonomiklik açısından 4 puanlayıcı ve 20 maddenin kullanılmasının güvenilir bir ölçme işlemi için bir sakınca oluşturmayacağı söylenebilir.

Karar çalışması neticesinde elde edilen iki farklı desene ait güvenirlik değerleri incelendiğinde çaprazlanmış desenin (G=0.96, Φ=0.45) kısmi yuvalanmış desene (G=0.92, Φ=0.27) göre daha güvenilir sonuçlar vermesi dikkat çekici bir noktadır. Bu sonuç, Doğan ve Anadol (2016)'un çalışmalarındaki çapraz desene ait G katsayısını 0.88, phi katsayısını 0.80, yuvalanmış desene ait G katsayısını 0.80, phi katsayısını 0.71, Taştan ve Uzun (2021)'un çalışmalarındaki çapraz desene ait G katsayısını 0.84, phi katsayısını 0.78, yuvalanmış desene ait G katsayısını 0.75, phi katsayısını 0.56 olarak elde ettikleri bulgularla paralellik göstermekte iken, Lee ve Cha (2016)'nın çalışmalarındaki çapraz desene ait G katsayısını 0.58, phi katsayısını 0.49, yuvalanmış desene ait G katsayısını 0.59, phi katsayısını 0.51; Yılmaz ve Gelbal (2011)'ın çalışmalarındaki çapraz desene ait G katsayısını 0.91, phi katsayısını 0.90, yuvalanmış desene ait G katsayısını 0.94, phi katsayısını 0.93 olarak buldukları bulgular ile ters düşmektedir. Ayrıca alanyazındaki bu çalışmalarda genellikle G ve Phi katsayılarının güvenilir seviyede çıkarken, G katsayıları Phi katsayılarına göre bir miktar yüksek değerde bulunmuştur. Nitekim Atılgan (2019), Phi katsayısının mutlak hataları, G katsayısının ise yalnızca bağıl hataları dikkate aldığından, Phi katsayısının genellikle G

katsayından daha düşük çıktığını belirtmiştir. Bu çalışma da yer alan her iki desenin de bağıl değerlendirme açısından yüksek güvenirlik gösterirken mutlak değerlendirmede yetersiz kalmasının sebebi, az sayıda projenin çalışma sürecine dahil edilmesi olabilir. Shavelson ve Webb (1991)'de özellikle küçük örneklemlerde mutlak değerlendirmenin daha düşük güvenirlik gösterdiğini vurgulamışlardır.

Özetle araştırmadaki iki farklı desen simülasyonu ile ulaşılan sonuçlar, tüm puanlayıcıların tüm kriterleri (maddeleri) değerlendirmesinin, dereceli puanlama anahtarındaki maddelerin uzmanlık alanlarına göre bölünmesine kıyasla daha tutarlı ve genellenebilir sonuçlar verdiğini göstermektedir. Benzer bir çalışma olan Lane ve Stone'un (2006) performans değerlendirme süreçlerinde bütüncül değerlendirme yaklaşımlarının daha tutarlı sonuçlar ortaya koyduğunu belirtmeleri bu sonucu desteklemektedir. Uygulanan her iki desen için mutlak değerlendirmenin güvenilir olmadığı, araştırmada kullanılan ölçme aracı ile genellenebilir sonuçlar elde edebilmek için bağıl değerlendirmenin kullanılması gerektiği sonucuna ulaşılmıştır.

Sınırlılıklar

Bu araştırmanın bazı sınırlılıkları bulunmaktadır:

1. Bu araştırma kontrollü şekilde oluşturulan simüle bir veri seti ile gerçekleştirilmiştir. Gerçek proje yarışmalarında ortaya çıkabilecek puanlayıcı yanlılıkları, değerlendirme yorgunluğu ve bağlamsal faktörler simülasyonda tam olarak yansıtılamamış olabilir.

2. Çalışma, 10 proje ile sınırlıdır. Daha büyük örneklemlerle yapılacak çalışmalar, bulguların genellenebilirliğini artırabilir.

3. Simülasyon, belirli bir proje türü (robotik-kodlama) ve değerlendirme aracı (TÜBİTAK formu) temel alınarak tasarlanmıştır. Farklı alan ve formlarla yapılacak çalışmalar bulgulara zenginlik katabilir.

Bu sınırlılıklara rağmen, araştırma bulguları proje yarışmalarının değerlendirme süreçlerinin tasarlanmasında önemli metodolojik katkılar sunmaktadır.

Öneriler

1. Bu çalışmada elde edilen bulgular, gerçek proje yarışması verileri ile test edilebilir. Ayrıca simülasyon sonuçları ile gerçek veri sonuçları karşılaştırılarak simülasyonun geçerliği değerlendirilebilir.

2. Farklı proje türleri (sosyal bilimler, fen bilimleri, sanat projeleri) ve farklı değerlendirme araçlarıyla benzer çalışmalar gerçekleştirilebilir.

3. Puanlayıcı eğitiminin değerlendirme güvenirliğine etkisini inceleyen deneysel çalışmalar yapılabilir.

4. Daha büyük örneklemlerle (20+ proje) gerçekleştirilecek çalışmalar, bulguların genellenebilirliğini artırabilir.

Çıkar Çatışması ve Etik Beyanı

Yazar herhangi bir çıkar çatışması beyan etmemektedir. Bu araştırma çalışması, araştırma yayın etiğine uygundur. IStL'de yayımlanan yazıların bilimsel ve hukuki sorumluluğu yazarlara aittir.

Yazarlık Katkı Beyanı

Yazar 1: Araştırma, Kaynaklar, Görselleştirme, Yazılım, Biçimsel Analiz ve Yazım-orijinal taslak.

KAYNAKÇA

Aktaş, M., & Alıcı, D. (2017). Kontrol Listesi, Analitik Rubrik ve Dereceleme Ölçeklerinde Puanlayıcı Güvenirliğinin Genellenebilirlik Kuramına Göre İncelenmesi. International Journal of Eurasia Social Sciences (Uluslararası Avrasya Sosyal Bilimler Dergisi), 8(29), 991-1010.

Atılgan, H. (2019). Genellenebilirlik Kuramı ve Uygulaması (1. Baskı). Anı Yayıncılık.

Ayvaz Tuncel, Z. (2021). Proje Tabanlı Öğrenme. Eğitimde Proje Geliştirme ve Yönetme içinde (Ed: Murat Taşdan, Halil İbrahim Kaya). Pegem Akademi Yayıncılık. 9786257676823

Baykul, Y. (2021).Eğitimde ve Psikolojide Ölçme: Klâsik Test Teorisi ve Uygulaması(4. Baskı). Ankara: Pegem Akademi Yayıncılık.

Bell, S. (2010) Project-Based Learning for the 21st Century: Skills for the Future, The Clearing House, 83(2), 39-43, DOI: 10.1080/00098650903505415

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

Büyükkıdık, S., & Anıl, D. (2015). Performansa Dayalı Durum Belirlemede Güvenirliğin Genellenebilirlik Kuramında Farklı Desenlerle İncelenmesi. Eğitim ve Bilim Dergisi, 40(177), 285-296.

Crocker, L., & Algina, J. (2006). Introduction to Classical and Modern Test Theory. Thomson Learning.

Chang, C. C., & Tseng, K. H. (2011). Using a Web-based portfolio assessment system to elevate project-based learning performances. Interactive Learning Environments, 19(3), 211-230.

Çetin, O. ve Şengezer, B. (2013). Ortaokul öğrencilerinin proje çalışmalarına ilişkin görüşleri. Ege Eğitim Dergisi, 14 (1), 24–49.

Eser, Ç. D., & Gelbal, S. (2012). Genellenebilirlik kuramı ve lojistik regresyona dayalı hesaplanan puanlayıcılar arası tutarlılığın karşılaştırılması. Kastamonu Üniversitesi Eğitim Fakültesi Dergisi, 21(2), 421-438.

Sun, A., Valiga, M. J., & Gao, X. (1997). Using Generalizability Theory to Assess the Reliability of Student Ratings of Academic Advising. The Journal of Experimental Education, 65(4), 367–379. http://www.jstor.org/stable/20152537

Göloğlu Demir, C. (2019). Öğretmenlerin TÜBİTAK ortaokul ve lise öğrencileri araştırma projeleri yarışması ve proje hazırlama sürecine yönelik görüşleri. II. Uluslararası İnsan ve Toplum Bilimleri Araştırmaları Kongresi Bildiri Kitabı, 4-6.

Güler, N. (2008). Klasik test kuramı, genellenebilirlik kuramı ve rasch modeli üzerine bir araştırma. (Doktora Tezi). Hacettepe Üniversitesi, Ankara.

Güler, N. (2009). Genellenebilirlik kuramı ve SPSS ile GENOVA programlarıyla hesaplanan G ve K çalışmalarına ilişkin sonuçların karşılaştırılması. Eğitim ve Bilim, 34(154), 93-103.

Gülle, A., Uzun, N. B., & Akay, C. (2018). Ortaokul Öğrencilerine Yönelik Blok Flüt İcra Performansı Dereceli Puanlama Anahtarının Güvenirliğinin Genellenebilirlik Kuramı ile İncelenmesi. İlköğretim Online, 17(3), 1463-1475.

Lafave, M. R. and Butterwick, D. J. (2014) A generalizability theory study of athletic taping using the technical skill assessment instrument. Journal of Athletic Training, 49(3), 368-372. https://doi.org/10.4085/1062-6050-49.2.22

Lane, S., & Stone, C. A. (2006). Performance assessment. In R. L. Brennan (Ed.), Educational measurement (4th ed., pp. 387-431). American Council on Education.

Lee, M., & Cha, D. (2016). A comparison of generalizability theory and many facet rasch measurement in an analysis of mathematics creative problem solving test. Journal of Curriculum Evaluation, 19(2), 251-279.

Menéndez-Varela, J. L., & Gregori-Giralt, E. (2018). The reliability and sources of error of using rubrics-based assessment for student projects. Assessment & Evaluation in Higher Education, 43(3), 488-499.

Millî Eğitim Bakanlığı [MEB] (2020). Araştırma ve Uygulamalarıyla Proje Temelli Öğrenme. Yenilik ve Eğitim Teknolojileri Genel Müdürlüğü, Ankara, Türkiye. Erişim: http://fclturkiye.eba.gov.tr/2020/09/07/arastirma-ve-uygulamalariyla-proje-tabanli-ogrenme

Millî Eğitim Bakanlığı [MEB] (2024). 16. Uluslararası MEB Robot Yarışması Uygulama Kılavuzu. Mesleki ve Teknik Eğitim Genel Müdürlüğü. https://robot.meb.gov.tr/organizasyon /uygulama-kilavuzu

Özbaşı, D. & Arcagök, S. (2021). Öğrenci projelerinin genellenebilirlik kuramı ile incelenmesi. Eğitimde Kuram ve Uygulama, 17(2), 69-78. doi: 10.17244/eku.1024532

Price, L.R. (2021). Psikometrik Yöntemler, Kuramdan Uygulamaya. Çeviri Editörü: Arif Özer, Burcu Atar. Mentis Yayıncılık.

Schwab, K. (2016). The Fourth Industrial Revolution. Crown Business.

Shavelson, R. J., & Webb, N. M. (1991). Generalizability Theory: A Primer. SAGE Publications.

Taştan, Z., & Bilge Uzun, N. (2021). Genellenebilirlik Kuramında Çok Yüzeyli Desenlerin İncelenmesi. Türkiye Sosyal Araştırmalar Dergisi, 25(3), 743-756.

Türkiye Bilimsel ve Teknolojik Araştırma Kurumu (TÜBİTAK). (2019). 2242-Üniversite Öğrencileri Araştırma Proje Yarışmaları Proje Rehberi. https://www.usak.edu.tr/UsersData/duyuru /1883/uni_proje_rehberi.pdf

Türkiye Bilimsel ve Teknolojik Araştırma Kurumu (TÜBİTAK). (2024). 2204-B Ortaokul Öğrencileri Araştırma Projeleri Yarışması Proje Rehberi. https://tubitak.gov.tr/sites/default/files/2024-10/ortaokul_proje_rehberi_2024-2025.pdf

Urbina, S. (2004). Essentials of Psychological Testing. John Wiley & Sons.

VanLeeuwen, D. M. (1997). Assessing Reliability Of Measurements With Generalizability Theory: An Application To Inter-Rater Reliability. Journal of Agricultural Education, 38(3), 36–42. https://doi.org/10.5032/jae.1997.03036

Yılmaz, F. N., & Gelbal, S. (2011). İletişim becerileri istasyonu örneğinde genellenebilirlik kuramıyla farklı desenlerin karşılaştırılması. Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 41(41), 509-518.

Yin, Y., & Shavelson, R. J. (2008). Application of Generalizability Theory to Concept Map Assessment Research. Applied Measurement in Education, 21(3), 273–291. https://doi.org/10.1080/08957340802161840.

Extended Abstract

Background and Purpose:

Project-based learning has become an essential pedagogical approach for cultivating individuals who produce, design, and innovate rather than consume. National and international competitions such as TUBİTAK, Teknofest, and Ministry of Education project contests support this pedagogical model by encouraging students to conduct research, design technological solutions, and integrate interdisciplinary knowledge into authentic problem-solving tasks. However, the increasing number of applications and the subjective structure of jury-based evaluations raise concerns regarding the reliability of scoring processes. Raters in such competitions differ in expertise, scoring behavior, and judgment tendencies, which may introduce substantial measurement error. Generalizability Theory (G-Theory) provides an advanced methodological framework to analyze multiple sources of measurement error and to design more reliable assessment systems.

This methodological simulation study aims to compare two different scoring designs based on G-Theory—crossed and partially nested designs—commonly encountered in project competitions. Specifically, the study investigates how different rater structures affect reliability indices (G and Phi coefficients) and variance components, and it explores optimal configurations of raters and items for improving evaluation reliability.

Method:

The study employed a simulated dataset created using RStudio to emulate realistic scoring patterns observed in robotics and coding project competitions. Ten projects were generated, scored by two groups of raters:

(1) Three project-writing experts, and

(2) Three information technologies experts.

Scores were produced to reflect variation across projects, raters, items, and associated interaction effects. A 20-item rubric based on the official TÜBİTAK 2204-B Technological Design evaluation form was used. Two distinct G-Theory designs were modeled:

Crossed design (b × m × p):

All three raters scored all 10 projects on all 20 items, yielding a fully crossed structure.

Partially nested design (b × (p : m)):

Items were split based on domain expertise.

Project-writing experts scored 7 items.

IT experts scored 13 items.

Six raters in total scored each project, but each rater evaluated only items relevant to their specialty.

For each design, variance components were estimated using ANOVA, followed by G-Study and D-Study analyses. Reliability indices (G and Phi) were produced under varying numbers of items and raters to determine optimal scoring configurations.

Results:

Across both designs, the variance attributable to the projects themselves was low (6.9%–8.1%), indicating limited differentiation among projects based on the simulated scores. This aligns with prior literature showing project-based assessments often generate low person variance due to similarities in student performance or rubric limitations.

The rater variance was substantial in both designs:

Crossed: 25.3%

Nested: 31.4%

This confirms that raters differ meaningfully in severity and leniency, reinforcing the necessity of rater training or improved scoring protocols. Interaction components—particularly the residual error term—accounted for the largest source of variability (31%–34%), suggesting the presence of uncontrolled measurement noise.

Regarding reliability indices:

Crossed design produced higher reliability ($G = 0.96$; $\Phi = 0.45$).

Nested design produced lower reliability ($G = 0.92$; $\Phi = 0.27$).

Both designs achieved high reliability for relative decisions, yet neither reached acceptable levels for absolute decisions, consistent with literature emphasizing that absolute decisions require more items, more raters, or less item variability.

Decision studies showed that increasing the number of raters and items boosted both G and Phi coefficients. For instance, the crossed design achieved acceptable precision with as few as 2 raters and 10 items, while the nested design required 4 or more raters to attain comparable reliability levels. These results indicate that distributing items across specialist raters increases measurement error relative to having all raters evaluate all items.

Discussion:

Findings demonstrate that although both scoring structures yield adequate reliability for relative decisions, the crossed design is consistently superior. Allowing all raters to score all criteria minimizes unwanted variability introduced when items are split across specialists. The nested design results indicate that domain-specific item assignment—though common in practice—may inadvertently amplify rater variability and reduce reliability.

The high residual variance in both designs reflects challenges inherent in project evaluation, including rubric interpretation differences, item difficulty discrepancies, and scorer subjectivity. These results support recommendations from previous studies emphasizing the importance of rater training, holistic rubric design, and consistency checks.

Conclusion and Implications:

This study contributes important methodological insights for organizers of project competitions. The findings indicate that:

Fully crossed scoring designs provide more reliable measurement than item-specialized rater assignments.

Increasing the number of raters enhances reliability, but in crossed designs, fewer raters may still be adequate.

Absolute decisions require more stringent scoring designs; hence relative decisions are more appropriate under typical competition conditions.

Organizers should prioritize scoring processes where all raters evaluate all items when feasible, apply systematic rater calibration, and use G-Theory to inform rubric design and evaluator allocation.

Limitations:

The simulation reflects only robotics/coding competitions and relies on a limited project pool. Future studies should replicate the methodology using real contest data, larger samples, and different disciplines.