

FEN BİLİMLERİ ÖĞRETMEN ve ÖĞRETMEN ADAYLARINA YÖNELİK PERFORMANS DEĞERLENDİRME RUBRİĞİ GELİŞTİRİLMESİ

DEVELOPING A PERFORMANCE ASSESSMENT RUBRIC FOR EVALUATING THE PROFESSIONAL COMPETENCIES OF SCIENCE TEACHERS AND TEACHER CANDIDATES

Nuri Mustafa SÜRÜCÜ

Fen Bilimleri Öğretmeni, Milli Eğitim Bakanlığı, İzmir, Türkiye

ORCID: <https://orcid.org/0000-0001-6273-7911>

nmsurucu07@gmail.com

Mehmet ŞAHİN

Prof.Dr., Dokuz Eylül Üniversitesi, Buca-İzmir, Türkiye

ORCID: <https://orcid.org/0000-0003-4247-483X>

mehmet.sahin@deu.edu.tr

Hüseyin Cihan BOZDAĞ

Fen Bilimleri Öğretmeni, Milli Eğitim Bakanlığı, İzmir, Türkiye

ORCID: <https://orcid.org/0000-0001-6735-7096>

chnbzd@gmail.com

Received: May 9, 2025

Accepted: July 7, 2025

Published: July 31, 2025

Suggested Citation:

Sürücü, N. M., Şahin, M., & Bozdağ, H. C. (2025). Fen bilimleri öğretmen ve öğretmen adaylarına yönelik performans değerlendirme rubriği geliştirilmesi. *International Journal of New Trends in Arts, Sports & Science Education (IJTASE)*, 14(3), 77-91.



Copyright © 2025 by author(s). This is an open access article under the [CC BY 4.0 license](https://creativecommons.org/licenses/by/4.0/).

Öz

Bu çalışma, fen bilimleri öğretmen ve öğretmen adaylarının mesleki yeterliklerinin ulusal ve uluslararası standartlara uygun biçimde değerlendirilebilmesi amacıyla geçerli ve güvenilir bir performans değerlendirme rubriği geliştirmeyi hedeflemiştir. Tarama yöntemiyle yürütülen çalışmada, rubrik; doküman analizi, uzman görüşleri ve ardışık pilot uygulamalar yoluyla oluşturulmuştur. Madde havuzu, ulusal öğretmenlik yeterlikleri ile uluslararası rubrik örneklerinden türetilmiştir. Kapsam geçerliği, Lawshe yöntemiyle uzman değerlendirmesi alınarak sağlanmış; yapı geçerliği ise korelasyon ve KMO analizleriyle test edilmiştir. Güvenirlilik Cronbach Alfa, Cohen Kappa ve sınıf içi korelasyon katsayıları ile değerlendirilmiştir. Nihai rubrik, yüksek iç tutarlılık ($\alpha = .968$) ve puanlayıcılar arası yüksek uyum ($ICC = .92-.96$) sergileyen 14 maddeden oluşmaktadır. Pilot uygulama bulgularına göre sınıf ortamında gözlemlenemeyen maddeler elenmiştir. Geliştirilen rubrik, fen eğitiminde öğretmen yeterliklerinin çok boyutlu, nesnel ve psikometrik açıdan güçlü biçimde değerlendirilmesine olanak sunmaktadır.

Anahtar Kelimeler: Performans değerlendirme, rubrik, fen eğitimi, geçerlik, güvenilirlik.

Abstract

This study aimed to develop a valid and reliable performance assessment rubric for science teachers and teacher candidates in accordance with national and international standards. Employing a descriptive research design, the rubric was constructed through document analysis, expert reviews, and iterative pilot testing. The item pool was derived from national teaching competencies and internationally recognized rubrics. Content validity was established via expert evaluation using the Lawshe method; construct validity was assessed through correlation and KMO analyses. Reliability was evaluated using Cronbach's alpha, Cohen's Kappa, and intraclass correlation coefficients. The final rubric consists of 14 items, each demonstrating strong internal consistency ($\alpha = .968$) and high inter-rater reliability ($ICC = .92-.96$). Items lacking observable classroom performance were excluded based on pilot findings. The developed rubric provides a comprehensive, objective, and psychometrically sound tool for evaluating multifaceted teaching competencies in science education. This instrument addresses the need for standardized performance assessment in teacher education and professional development processes.

Keywords: Performans assesment, rubric, science education, validity, reliability.

GİRİŞ

Geçmişten günümüze doğru bilgi ve teknolojiadaki sürekli değişim ve yenilikler, bunlara uyum sağlayabilecek nitelikli insanların ihtiyacını ortaya çıkarmaktadır. Nitelikli insan özelliklerinin bireyler tarafından kazanılmasında eğitim kurumları ve bu kurumda çalışan öğretmenlerin niteliği önem kazanmaktadır (Darling-Hammond, 2000) Öğrencilerin 21. Yüzyıl becerilerinden; eleştirel düşünme, problem çözme, iletişim, işbirliği, yaratıcılık ve yenilik gibi yetkinlikleri geliştirebilmeleri için bilimsel süreç becerilerine hakim olmaları gerekmektedir (Petrie, 2023). Fen bilimleri eğitimi, öğrencilere bilimsel kavramları öğretmenin yanı sıra onların bilimsel düşünme, problem çözme ve analitik becerilerini geliştirmeyi hedefleyen disiplinlerarası bir süreçtir. Aynı zamanda fen bilimleri eğitimi bireylerin bilimsel okuryazarlık kazanmasını, analitik düşünme becerilerini geliştirmesini ve bilimsel süreçleri anlamasını sağlamayı hedefleyen bir öğretim alanıdır (National Research Council-NRC, 2012). Bu süreçte öğretmenlerin pedagojik ve alan bilgisi yeterlilikleri, öğrenci başarısını doğrudan etkileyen kritik faktörler arasında yer almaktadır (Shulman, 1986). Ancak fen bilimleri öğretmenlerinin sınıf içi performanslarını objektif kriterler doğrultusunda değerlendirmek, eğitimde nitelik güvencesi sağlamak açısından önemli bir gerekliliktir; öğretmenlerin sahip olduğu nitelikler ile öğrenci başarısı arasında bir ilişki olduğu da bilinmektedir (Darling-Hammond, 2000). Öğretmenlerin sınıf içi uygulamalarını ve pedagojik becerilerini değerlendirmek için çeşitli yöntemler geliştirilmiştir. Bunlar arasında ders gözlemleri, öğrenci geri bildirimleri, öz değerlendirme formları ve rubrikler yer almaktadır (Marzano, 2012). Rubrikler önceden belirlenmiş ölçütler çerçevesinde, öğretmen performansını değerlendiren, farklı seviyelerde puanlama imkanı sunan ve öğretmenlere gelişim alanlarını gösteren bir değerlendirme aracıdır (Brookhart, 2013). Yapılan eğitim araştırmaları, rubriklerin öğretmen değerlendirme süreçlerinde formatif (süreç odaklı) ve sumatif (sonuç odaklı) değerlendirme amacıyla etkili şekilde kullanılabileceğini göstermektedir (Jonsson & Svingby, 2007).

Günümüzde fen bilimleri öğretmen ve öğretmen adaylarının mesleki performansının çok boyutlu, nesnel ve güvenilir biçimde değerlendirilmesi kritik bir öneme sahiptir. Bu sayede hem öğretmen yetiştirme programlarının hem de çağdaş eğitim standartlarının artırılması sağlanabilir (Chasteen & Scherr, 2020; Edwards, 2017). Performans değerlendirme süreçlerinde, öğretmenlerin yalnızca alan bilgisini değil; pedagojik uygulama, iletişim, teknoloji entegrasyonu, disiplinlerarası düşünme ve materyal geliştirme gibi karmaşık becerilerini kapsayan bir yeterlikler bütününe odaklanılması gerektiği, güncel araştırmalarda sıklıkla vurgulanmaktadır (Aktaş & Özmen, 2022; Santiani vd., 2023; Syawaludin vd., 2024).

Öğretmen performansının sistematik, şeffaf ve adil biçimde değerlendirilmesinde, dereceli puanlama anahtarı şeklinde yapılandırılan rubriklerden yaygın olarak yararlanılmaktadır. Rubrikler, belirli bir beceri ya da performans alanında, önceden tanımlanmış ölçütler ve düzeyler aracılığıyla gözlemlenen davranışların objektif ve dereceli biçimde puanlanmasını sağlayan yapılandırılmış ölçme araçlarıdır (Brookhart, 2013; Jonsson & Svingby, 2007). Bu kapsamda geliştirilen rubrikler, öğretmen performansının sistematik ve güvenilir biçimde değerlendirilmesine olanak tanımaktadır. Alanyazında değerlendirme rubrikleri, fen bilgisi/fen bilimleri eğitimi, biyoloji, fizik, kimya, laboratuvar uygulamaları, hemşirelik, disiplinlerarası düşünme, deneysel tasarım, teknoloji entegrasyonu ve iletişim/öğretim becerileri başta olmak üzere çeşitli disiplinlerde etkin biçimde kullanılmaktadır (Aktaş & Özmen, 2022; Alias & Osman, 2015; Amelia et al., 2024; Antony & Paidi, 2019; Aumann et al., 2023; Bartels et al., 2019; Huda et al., 2020; Permana & Widodo, 2022; Santiani et al., 2023; Yeşildağ-Hasancebi et al., 2022). Nitekim, Aumann, Schnebel ve Weitzel (2023) tarafından geliştirilen EnTPACK rubriği, teknolojik pedagojik alan bilgisi yeterliklerini bütüncül olarak değerlendirirken; uzman görüşü, Cohen's Kappa (.704) ve ICC (.873) ile yüksek kapsam geçerliği ve puanlayıcılar arası güvenilirlik sunmuştur. Aktaş ve Özmen (2022) ise fen bilgisi öğretmen adaylarıyla yürüttükleri çalışmada, kapsam geçerliğini uzman görüşüyle; iç tutarlılığı (.942) ve puanlayıcılar arası uyumu (.962) yüksek değerlerle raporlamıştır. Benzer şekilde, Alias ve Osman (2015), Amelia vd. (2024) ve Antony ve Paidi (2019), farklı ülkelerde geliştirdikleri rubriklerde kapsam ve yapı geçerliğini detaylı biçimde analiz etmiş; yüksek Cronbach Alpha ve ICC gibi istatistiksel göstergelerle araçlarının güvenilirliğini desteklemiştir.

Santiani vd. (2023), disiplinlerarası düşünme becerilerine odaklanan IITSSL rubriğinde tüm boyutlar için ICC'nin .637 olarak bulunduğunu bildirmiştir. Permana ve Widodo (2022) ise Endonezya'da geliştirdikleri rubrikte üç uzman görüşüyle içerik geçerliğini sağlamış; ICC'yi .96 ve maddelerin %84'ünde tam uyumu raporlamıştır. Türkiye'de Yeşildağ-Hasancebi vd. (2022), kapsam geçerliğinde CVR ($> .80$), CVI (.94), puanlayıcılar arası güvenirlikte ise Cohen's Kappa (.60-.91) ve Pearson korelasyon ($r = .96$) ile yüksek değerler elde etmişlerdir. Disher (2018), pilot uygulama sürecinde pratikte gözlemlenemeyen maddeleri rubrikten çıkarmış ve böylece ölçme aracının işlevselliğini güçlendirmiştir.

Bunlara ek olarak, Bartels vd. (2019) Almanya ve Avustralya'da geliştirilen performans odaklı rubrikte uzman paneliyle yüksek kapsam ve yapı geçerliği, iç tutarlılık ($\alpha = .56-.79$) ve Cohen's Kappa (.78) gibi göstergelerle güçlü psikometrik özellikler elde etmişlerdir. Huda et al. (2020) tarafından geliştirilen laboratuvar uygulama rubriğinde; madde-toplam korelasyonlarının .468'in üzerinde olduğu, Cronbach's alpha katsayısının .90 olarak belirlendiği ve çoklu puanlayıcılar arası güvenirliğin yüksek düzeyde gerçekleştiği belirtilmiştir; ayrıca uzman görüşüyle kapsam geçerliği sağlanmıştır.

Bu zengin ulusal ve uluslararası literatür, öğretmen performans değerlendirmesinde nesnel ve güvenilir ölçme araçlarının geliştirilmesinin önemini ortaya koysa da mevcut çalışmaların önemli bir kısmı Türkiye'nin güncel ulusal standartları ve uygulama gereksinimleriyle tam uyumlu değildir. Özellikle Millî Eğitim Bakanlığı'nın son yıllarda öğretmen alım süreçlerinde bransa özgü performans değerlendirmelerine ve mülakat uygulamalarına ağırlık vermesi, bu alanda geçerli ve güvenilir, alan odaklı yeni araçlara olan gereksinimi artırmaktadır (MEB, 2023).

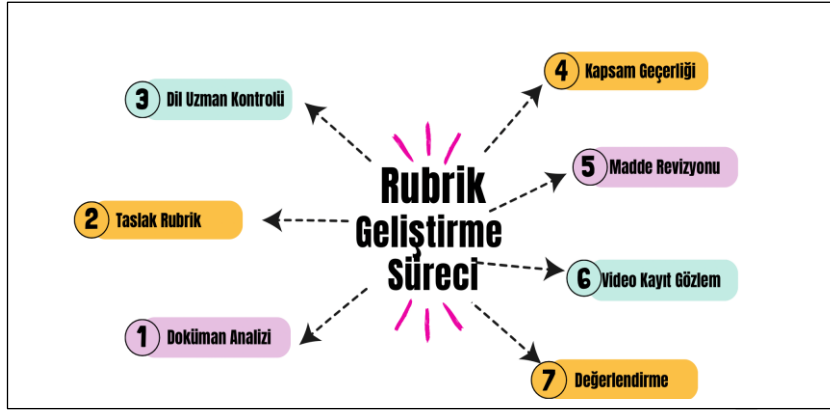
Aynı şekilde, öğretmen adaylarının lisans eğitiminin son yılında yürütülen öğretmenlik uygulaması dersi kapsamında, danışman öğretim üyeleri ve uygulama okulu öğretmenleri tarafından yapılan performans puanlamalarında da nesnel, standardize edilmiş ölçütlerin eksikliği önemli bir sorun oluşturmaktadır. Literatürde, mevcut değerlendirmelerin çoğunlukla öznel yargılara dayandığı ve bu nedenle değerlendirmelerin güvenirliğinin sınırlı olduğu vurgulanmaktadır. Bu açıdan, standart ve güvenilir bir performans değerlendirme rubriğinin geliştirilmesi hem eğitimde kalite güvencesini hem de değerlendirme süreçlerinde şeffaflık ve nesnelliği güçlendirecektir (Aktaş & Özmen, 2022).

Bu çalışma, Millî Eğitim Bakanlığı tarafından belirlenen öğretmen yeterlikleri ve uluslararası standartlar temelinde, fen bilimleri öğretmenleri ve öğretmen adaylarının performansını nesnel, geçerli ve güvenilir biçimde değerlendirecek bir rubrik geliştirmeyi amaçlamaktadır. Geliştirilen rubrik, kapsam geçerliği, yapı geçerliği, iç tutarlılık ve puanlayıcılar arası güvenirlik analizleriyle psikometrik olarak test edilmiş ve alana, ulusal standartlarla uyumlu yeni bir değerlendirme aracı kazandırılması amaçlanmaktadır.

YÖNTEM

Bu araştırma, fen bilimleri öğretmen ve öğretmen adaylarının mesleki performanslarının geçerli ve güvenilir biçimde ölçülmesini sağlayacak bir değerlendirme rubriği geliştirmek amacıyla, tarama yöntemi araştırma kullanılarak gerçekleştirilmiştir. Yöntemsel araştırmalar, ölçme araçlarının geliştirilmesi ve psikometrik niteliklerinin sistematik olarak incelenmesini amaçlar (Fraenkel, Wallen & Hyun, 2012). Bu doğrultuda, çalışmada fen bilimleri öğretmen ve öğretmen adaylarının çok boyutlu, nesnel ve güvenilir şekilde değerlendirilmesini mümkün kılacak bir rubriğin geliştirilmesi hedeflenmiştir.

Çalışma kapsamında fen bilimleri öğretmen ve öğretmen adaylarının performans değerlendirilmesi belirlenen ölçütler aracılığıyla gözlem yapılması ve bu gözlemlerin puanlanmasına imkan tanıyan rubrik geliştirilerek sağlanmıştır. Değerlendirme rubriği şeklinde hazırlanan analiz aracının geliştirilmesi sürecinin tüm aşamalarında araştırmacı bizzat rol almış olup; rubrik hazırlanmasına ilişkin tüm işlemler aşamalar halinde rubrik geliştirme süreci şeklinde sunulmuştur.



Şekil 1. Performans değerlendirme rubriği geliştirme süreci

Şekil 1'e bakıldığında, fen bilimleri öğretmenlerine yönelik performans değerlendirme rubriği geliştirme sürecinin sırasıyla; (1) doküman analizi, (2) taslak rubrik oluşturma, (3) dil uzmanı kontrolü, (4) kapsam geçerliği çalışması, (5) madde revizyonu, (6) video gözlem ve (7) değerlendirme olmak üzere yedi ardışık aşamadan oluştuğu anlaşılmaktadır.

2.1. Katılımcılar ve Uzman Seçimi

Kapsam geçerliği çalışmalarında mümkün olduğunca fazla sayıda ve farklı kurumlardan uzman görüşüne başvurulması önerilmektedir (DeVellis, 2017). Uzmanların seçiminde fen eğitimi alanındaki akademik yetkinlik ve ilgili çalışma deneyimi temel alınmıştır. Bu davete 12 öğretim üyesi olumlu yanıt vermiş, kapsam geçerliği formunu ise 3 öğretim üyesi eksiksiz doldurmuştur. Ayrıca İzmir'de çeşitli devlet okullarında görev yapan 10 fen bilimleri öğretmeni de uzman gruba dahil edilmiştir.

tablo 1. geçerlik ve güvenirlik çalışmalarına katılan katılımcı grupları ve özellikleri.

Çalışma Türü	Katılımcı Grubu	Sayı	Kurum/Bölge
Geçerlik	Türkçe Öğretmeni	2	İzmir (devlet okulları)
Geçerlik	Akademisyenler	3	Ankara, İstanbul, İzmir
Geçerlik	Fen Bilimleri Öğretmeni	10	İzmir (devlet okulları)
Güvenirlik	Fen Bilimleri Öğretmeni	15	İzmir (Buca, Karşıyaka, Konak)
Güvenirlik	Öğretmen Adayı	11	İzmir
Güvenirlik	Puanlayıcı	2	İzmir (devlet okulları)

Katılımcılar amaçlı örnekleme ile seçilmiş; alan uzmanlığı, öğretmenlik deneyimi, staj sürecinde bulunma ve gönüllülük ölçütleri dikkate alınmıştır (Creswell, 2017; Yıldırım & Şimşek, 2022). Bununla birlikte, video kaydı gerekliliği ve her bir uygulamanın iki bağımsız puanlayıcı tarafından ayrıntılı biçimde değerlendirilmesinin yüksek zaman ve emek gerektirmesi nedeniyle, uygulayıcı örneklem büyüklüğü gönüllü katılım ve pratik sınırlılıklar çerçevesinde şekillenmiştir. Tüm bu zorluklara rağmen, veri toplama sürecinde İzmir'in üç farklı ilçesinden öğretmen ve öğretmen adaylarının katılımı sağlanarak, örnekleme bölgesel çeşitlilik gözetilmiştir. Puanlayıcı sayısı ise, ölçme değerlendirme literatüründe Cohen's Kappa analizinin yaygın olarak iki puanlayıcı ile yapılması nedeniyle ikiye sınırlandırılmıştır (Hallgren, 2012; Jonsson & Svingby, 2007).

Tablo 2. Araştırma katılımcılarının cinsiyet dağılımı.

Cinsiyet	Fen Bilimleri Öğretmeni	Fen Bilimleri Öğretmen Adayı	Toplam	Yüzde (%)
Kadın	10	9	19	73.1
Erkek	5	2	7	26.9
Toplam	15	11	26	100

2.2. Veri Toplama Süreci ve Araçları

Rubrik geliştirme sürecinde, öncelikle Millî Eğitim Bakanlığı'nın yayımladığı “Öğretmenlik Mesleği Genel ve Özel Alan Yeterlikleri” (MEB, 2008) ile “Öğretmenlik Mesleği Genel Yeterlikleri” (MEB, 2017) kitapçıkları sistematik doküman analizi yöntemiyle (Yıldırım & Şimşek, 2022) incelenmiştir. Ayrıca Australian Institute for Teaching and School Leadership (AITSL), Ministry of Education (MOE), National Board for Professional Teaching Standards (NBPTS), Office for Standards in Education, Children's Services and Skills (Ofsted) ve Performance Assessment for California Teachers (PACT) gibi uluslararası rubrikler analiz edilmiştir. Doküman analizi sonucunda elde edilen yeterlik alanlarına dayalı ölçülebilir davranışlar madde havuzunun temelini oluşturmuştur. Literatürde, kapsam geçerliğini artırmak için ölçek geliştirme sürecinde ölçme aracının temel aldığı kavramsal alanın çeşitli kaynaklardan ve değişik düzeylerden detaylı olarak incelenmesinin önemi vurgulanmaktadır. Bu doğrultuda, ulusal yasal düzenlemeler ve standartlar ile birlikte uluslararası örneklerin de dikkate alınarak madde havuzunun kapsamlı şekilde hazırlanması tavsiye edilmektedir (DeVellis, 2017).

Rubrik maddelerinin dört dereceli olarak tasarlanmasının nedeni hem doküman analizleri sonucunda hem de ulusal ve uluslararası literatürde (AITSL, 2023; MOE Singapore, 2022; NBPTS, 2022; Ofsted, 2019; PACT, 2018) dört dereceli sistemlerin yaygın ve etkili olduğunun görülmesidir. Dört dereceli sistem, puanlayıcıların kararsız (“orta”) seçeneğe yönelmesini engeller, davranış düzeylerini ayırıştırır ve puanlayıcılar arası tutarlılığı artırır (Brookhart, 2013; Jonsson&Svingby,2007; Stevens & Levi, 2013).

Rubrik maddelerinin dil ve anlatım açısından açık, anlaşılır ve hedef kitleye uygun olması amacıyla, iki yüksek lisans derecesine sahip iki Türkçe öğretmenine taslak rubrik matbu çıktılar halinde sunulmuştur. Dil uzmanları, her bir madde ve yönerge üzerinde tespit ettikleri dil bilgisi ve anlatım bozukluklarını kâğıt üzerinde işaretlemiş; araştırmacı, bu düzeltmeleri uygulayarak rubriği tekrar incelemeye sunmuştur. Böylece, ölçme aracının dilsel geçerliği ve anlaşılabilirliği artırılmış, maddelerin hedef kitle tarafından doğru anlaşılması sağlanmıştır.

2.3. Kapsam Geçerliği Süreci

Hazırlanan 19 maddelik taslak rubrik, kapsam geçerliği için 3 fen eğitimi öğretim üyesi ve 10 fen bilimleri öğretmenine yüz yüze ya da e-posta yoluyla sunulmuştur. Uzmanlara, (Lawshe tekniğine uygun olarak) her maddeyi “1 = Önemli değil”, “2 = Önemli”, “3 = Çok önemli” şeklinde üçlü Likert tipi derecelendirme ile puanlamaları ve her maddenin altına boş bırakılan alana madde ya da derecelendirme ile ilgili eleştiri/düzeltilme/öneri yazmaları istenmiştir. Kullanılan değerlendirme formu ve uzmanlara gönderilen bilgilendirici davet mektubu Ekler bölümünde sunulmuştur.

Uzmanlardan alınan nicel veriler, Lawshe (1975) yöntemiyle KGO (Kapsam Geçerliği Oranı) ve KGİ (Kapsam Geçerliği İndeksi) değerlerinin hesaplanmasında kullanılmıştır. KGO, her bir maddenin “gerekli” görülme oranını; KGİ ise tüm maddelerin ortalama geçerlik düzeyini gösterir (Ayre & Scally, 2014). Minimum anlamlı KGO değeri, uzman sayısına göre Ayre & Scally (2014) tablosundan alınmış ve 13 uzman için 0.54 olarak belirlenmiştir. Hesaplanan değerler Tablo 3'te özetlenmiştir.

Tablo 3. Maddelere ilişkin kapsam geçerlik oranları (KGO) ve kararları (örnek).

Madde No	Açıklama	Uzman Sayısı	“Önemli” Uzman (ne)	Minimum KGO	Gerçek KGO	Karar
1	Planlama	13	12	0.54	0.85	Kabul
2	Öğrenme Ortamı	13	10	0.54	0.54	Kabul
3	Materyal ve Kaynaklar	13	8	0.54	0.23	Reddedildi

Ayrıca uzmanların açık uçlu nitel önerileri doğrultusunda, madde ifadelerinde ve derecelendirmelerde düzenlemeler yapılmıştır. Lawshe analizi sonucunda performans gözleminin sistematik olarak güç olduğu belirlenen bir madde çıkarılmış; nitel dönütlerle yapılan düzeltmeler sonrasında madde sayısı 18’e indirilmiştir.

2.4. Pilot ve Ana Uygulama – Video Kayıt Gözlemi

Pilot uygulama sürecinde, iki bağımsız puanlayıcı, fen bilimleri öğretmen ve öğretmen adaylarından rastgele seçilen beş farklı dersin video kaydını, geliştirilen performans değerlendirme rubriğini kullanarak ayrı ayrı puanlamıştır. Değerlendirme sonrasında, puanlayıcılar arasında farklılık gösteren madde puanlamaları karşılıklı olarak görüşülerek ortak bir anlayışa varılmıştır. Benzer biçimde, Dishar (2018) da geliştirdikleri rubriğin uygulanabilirliğini ve maddelerin açıklığını pilot uygulama ile test etmiş; pilot aşama sonunda elde edilen bulgular doğrultusunda maddelerde gerekli revizyonları yapmışlardır. Puanlayıcılardan biri araştırmacı, diğeri ise doktora derecesine sahip bir fen bilimleri öğretmenidir. Araştırmacı, ikinci puanlayıcıya rubrik kullanımına yönelik eğitim sağlamış ve birlikte örnek videolar üzerinden uygulamalı değerlendirmeler gerçekleştirilmiştir (Brookhart, 2013; Stemler, 2004). Bu eğitim süreciyle puanlayıcılar arası kodlama tutarlılığı artırılmıştır.

2.5. Pilot Uygulama Sonrası Madde Revizyonları

Pilot değerlendirmeler sırasında bazı maddelerin sınıf ortamında uygulanabilirliğiyle ilgili önemli sınırlılıklar ortaya çıkmıştır. Özellikle Madde 4, 6, 13 ve 14 kapsamında, her iki puanlayıcı da video kayıtlarında bu maddelere ilişkin gözlemlenebilir herhangi bir performans bulgusuna rastlamadıklarını rapor etmiş ve ilgili maddeler için puanlama gerçekleştirememiştir. Bu durum, ilgili maddelerin sınıf ortamında sistematik ve nesnel olarak gözlemlenmesinin pratikte güç olduğuna işaret etmektedir. Pilot uygulama sonucunda, bazı maddelerin sınıf ortamında gözlemlenmesinin pratikte mümkün olmadığı belirlenmiş ve bu maddeler rubrikten çıkarılmıştır. Literatürde de gözlemlenebilirliği olmayan veya pratikte işlevsel olmayan maddelerin ölçekten çıkarılmasının hem uygulama geçerliğini hem de ölçümün güvenilirliğini artıracığı vurgulanmaktadır (Brookhart, 2013; DeVellis, 2017; Jonsson & Svingby, 2007). Aşağıdaki tabloda, söz konusu maddelerin neden gözlemde puanlanamadığına dair gerekçeler ayrıntılı olarak sunulmuştur.

Tablo 4. Pilot uygulama sonrası rubrikten çıkarılan maddeler ve gerekçeleri.

Madde No	Madde Adı	Kısa Açıklama	Çıkarılma Nedeni
4	Çevreyi inceleme ve tanıma merakı	Çevre gezileri, okul dışı etkinlikler ve kurumlarla iş birliği	Performans, ders saatinde doğrudan gözlemlenebilir ve uygulanabilir olmadığından rubrikten çıkarılmıştır.
6	Bilimin doğası ve tarihsel gelişimi	Bilimin doğası ve tarihsel gelişimiyle ilgili davranışlar	Tüm fen bilimleri kazanımlarında gözlemlenebilir olmaması nedeniyle rubrikten çıkarılmıştır.
13	Güvenlik önlemleri	Sınıf içi etkinliklerde güvenlik önlemlerinin uygulanması	Performansın doğrudan ve nesnel gözleminin mümkün olmaması nedeniyle rubrikten çıkarılmıştır.
14	Özel gereksinimli bireyler	Özel gereksinimli bireylerle ilgili uyarlamalar ve yaklaşımlar	Sınıf ortamında doğrudan ve güvenilir biçimde gözlemlenebilir olmaması; gözlemcinin bu öğrencileri ayırt edememesi nedeniyle rubrikten çıkarılmıştır.

“Çevreyi inceleme ve tanıma merakı” maddesi (Madde 4), çevre gezileri ve okul dışı etkinliklerin her ders saatinde doğrudan gözlemlenmesinin pratikte mümkün olmaması nedeniyle rubrikten çıkarılmıştır. Benzer şekilde, “Bilimin doğası ve tarihsel gelişimi” (Madde 6), “Güvenlik önlemleri” (Madde 13) ve

“Özel gereksinimli bireyler” (Madde 14) maddeleri de video gözlemlerinde doğrudan ve güvenilir biçimde değerlendirilememeleri; bu maddelerin fen bilimleri müfredatında yer alan her kazanım için uygun olamamaları nedeniyle rubrik kapsamı dışında bırakılmıştır. Nitekim, alanyazında da gözlemlenebilirliği ve ölçülebilirliği sınırlı olan maddelerin, rubrik veya ölçeklerden çıkarılmasının hem kapsam ve uygulama geçerliğini hem de puanlayıcılar arası güvenilirliği artırdığı vurgulanmaktadır (Brookhart, 2013; DeVellis, 2017; Jonsson & Svingby, 2007). Sonuç olarak, pilot uygulama ve gözlem verilerinden elde edilen bulgular doğrultusunda yapılan eleme ve revizyonlar sayesinde, rubrikte yalnızca sınıf ortamında doğrudan gözlemlenebilen ve güvenilir biçimde puanlanabilen maddelere yer verilmiştir.

2.6. Etik Onay ve Yasal İzinler

Araştırmanın tüm aşamaları için Dokuz Eylül Üniversitesi Etik Kurulu’ndan ve Millî Eğitim Bakanlığı’ndan gerekli izinler alınmış, tüm katılımcılardan yazılı onam sağlanmıştır. Araştırma, ulusal ve uluslararası etik standartlara uygun olarak yürütülmüştür (YÖK, 2021).

2.7. İstatistiksel Analizler

Kapsam geçerliği, Lawshe (1975) yöntemiyle hesaplanan KGO ve KGİ değerleri üzerinden değerlendirilmiştir (Ayre & Scally, 2014). Rubrik maddelerine ilişkin tanımlayıcı istatistikler (aritmetik ortalama, standart sapma, minimum ve maksimum değerler) hesaplanmıştır. Yapı geçerliği için Kaiser-Meyer-Olkin (KMO) örneklem yeterlilik testi ve Bartlett’s Sphericity testi uygulanmış; ek olarak madde arası korelasyon matrisi ve madde varyansları analiz edilmiştir (Field, 2018). Güvenirlik analizleri kapsamında, ölçeğin iç tutarlılığı için Cronbach’s Alpha katsayısı, puanlayıcılar arası tutarlılık için ise Cohen’s Kappa ve ICC (Intraclass Correlation Coefficient) analizlerinden yararlanılmıştır (Hallgren, 2012). Tüm analizler IBM SPSS 25 paket programı ile yürütülmüştür.

BULGULAR

3.1. Kapsam Geçerliği

Geliştirilen performans değerlendirme rubriğinin kapsam geçerliği, alan uzmanlarının görüşleri doğrultusunda Lawshe (1975) yöntemiyle değerlendirilmiştir. Her bir madde için elde edilen Kapsam Geçerlik Oranları (KGO), ölçme aracının her bir performans alanında yeterli düzeyde temsil gücüne sahip olup olmadığını ortaya koymaktadır. Aşağıda, rubrik maddelerine ilişkin kapsam geçerlik oranları sunulmuştur:

Tablo 5. Rubrik maddelerine ilişkin kapsam geçerlik oranları (KGO/) ve performans alanları.

Madde No	Performans Alanı	KGO	Madde No	Performans Alanı	KGO
1	Planlama	1.00	10	Bireysel Farklılıklar	1.00
2	Öğrenme Ortamının Planlanması	1.00	11	Bilimsel Gelişmeler ile Toplum Çevre Etkileşimi	0.54
3	Materyal ve Kaynaklar	1.00	12	Güvenlik Önlemleri	1.00
4	Çevreyi İnceleme ve Tanıma Merakı	0.69	13	Özel Gereksinimli Bireyler	0.69
5	Bilimsel Süreç Becerileri	1.00	14	Öğrenci Gelişimlerini İzleme	0.85
6	Bilimin Doğası ve Tarihsel Gelişimi	0.54	15	Dersin Sunulması	1.00
7	Eleştirel Düşünme ve Problem Çözme Becerisi	1.00	16	Sınıf Ortamı	0.85
8	Bilimsel ve Teknolojik Kavram Kullanımı	0.85	17	Ön Öğrenmelerle İlişki	0.54
9	Bilim ve Teknoloji İlişkisi	0.85	18	Atatürkçülük	-0.54
					KGİ
					0.83

Analiz sonucunda, toplam 18 maddeden 8’inin (Planlama, Öğrenme Ortamının Planlanması, Materyal ve Kaynaklar, Bilimsel Süreç Becerileri, Eleştirel Düşünme ve Problem Çözme Becerisi, Bireysel Farklılıklar ve Güvenlik Önlemleri) KGO değerinin 1.00 olduğu, yani tüm uzmanlar tarafından

“gerekli” olarak değerlendirildiği görülmüştür. Bu bulgu, rubriğin temel performans alanlarının uzmanlar arasında yüksek düzeyde görüş birliğiyle kabul edildiğini göstermektedir.

KGO değeri 1.00’ün biraz altında kalan maddeler ise, Madde 8 (Bilimsel ve Teknolojik Kavram Kullanımı) [0.85], Madde 9 (Bilim ve Teknoloji İlişkisi) [0.85], Madde 13 (Özel Gereksinimli Bireyler) [0.69], Madde 4 (Çevreyi İnceleme ve Tanıma Merakı) [0.69], Madde 6 (Bilimin Doğası ve Tarihsel Gelişimi) [0.54], Madde 11 (Bilimsel Gelişmeler ile Toplum Çevre Etkileşimi) [0.54] ve Madde 17 (Ön Öğrenmelerle İlişki) [0.54] olarak sıralanmıştır. Bu maddelerin de uzmanlar tarafından kabul edilebilir sınırdan değerlendirildiği görülmektedir. En düşük KGO değeri ise Atatürkçülük maddesinde (-0.54) bulunmuş ve bu madde, uzman görüşleri doğrultusunda rubrikten çıkarılmıştır. Madde sayısı 17’ye düşmüştür.

Rubriğin genel kapsam geçerlik indeksi (KGİ) 0.83 olarak hesaplanmıştır. Genel olarak, maddelerin büyük çoğunluğunun yüksek KGO değerlerine sahip olması, rubriğin kapsam geçerliğinin uzman görüşleriyle desteklendiğini göstermektedir.

3.2. Uzman Görüşlerine Dayalı Madde Revizyonu: Nitel Dönütlerin Rolü

Kapsam geçerliği sürecinde, 13 alan uzmanına iletilen değerlendirme formunda, her bir madde ve performans derecesinin altına açık uçlu geri bildirim yazmaları için boş alan bırakılmıştır. Bu yöntemle yalnızca Likert tipi nicel puanlar değil, aynı zamanda maddelerin kapsamı, açıklığı, uygulanabilirliği ve geliştirme önerilerine yönelik nitel dönütler de toplanmıştır.

“Eleştirel düşünme ve problem çözme becerisi” maddesine ilişkin birden fazla uzmandan alınan geri bildirimlerde, “Bu iki becerinin öğretmen performansında bağımsız olarak gözlemlenmesi gerektiği” ve “eleştirel düşünme ile problem çözme arasında işlevsel farklılıklar bulunduğu” belirtilmiştir. Bir uzman, ‘Bir öğretmen eleştirel düşünmeyi destekleyebilir ancak problem çözmede aynı yeterliliği gösteremeyebilir’ değerlendirmesinde bulunmuştur.

Alınan bu nitel dönütler doğrultusunda, söz konusu madde yeniden yapılandırılarak “eleştirel düşünme” ve “problem çözme” performans göstergeleri iki ayrı madde olarak rubriğe eklenmiştir. Böylece hem kapsam geçerliğinin hem de ölçme aracının hassasiyetinin artırılması hedeflenmiştir.

Kapsam geçerliği sürecinde, uzmanlardan alınan nitel dönütler doğrultusunda “Atatürkçülük” maddesinin performans gözlemine uygun bulunmaması nedeniyle rubrikten çıkarılmış ve madde sayısı 17’ye düşürülmüştür. Ancak, “eleştirel düşünme ve problem çözme becerisi” maddesine ilişkin alınan uzman geri bildirimleri sonucunda, bu iki yetkinlik ayrı ayrı maddeler olarak yapılandırılmıştır. Böylece, çıkarılan maddeye karşılık, bir maddenin ikiye ayrılmasıyla rubrikte toplam madde sayısı yeniden 18’e yükselmiştir. Bu süreç, hem kapsam geçerliği hem de ölçme aracının ayırt ediciliği açısından önemli bir iyileştirme olarak değerlendirilmiştir.

3.3. Pilot Uygulama Sonrası Madde Revizyonları

Pilot uygulama sürecinde elde edilen video gözlemleri ve puanlayıcı analizleri sonucunda, rubriğin sınıf ortamında sistematik ve nesnel biçimde uygulanabilirliğini sınırlayan maddeler belirlenmiş ve gerekli revizyonlar yapılmıştır. Bu kapsamda, puanlayıcılar tarafından gözlem sırasında puanlanamayan dört madde (Madde4: Çevreyi inceleme ve tanıma merakı, Madde6: Bilimin doğası ve tarihsel gelişimi, Madde13: Güvenlik önlemleri ve Madde14: Özel gereksinimli bireyler) rubrikten çıkarılmıştır. Ayrıca, kapsam geçerliği sürecinde uzman görüşlerinden elde edilen nitel dönütler doğrultusunda, “eleştirel düşünme ve problem çözme becerisi” maddesi iki ayrı performans göstergesi olarak yapılandırılmıştır. Böylece başlangıçta hazırlanan rubrik 18 maddeden oluşmaktaydı. Uzman görüşleriyle bir madde çıkarılmış ve “eleştirel düşünme ve problem çözme” maddesi iki ayrı performans göstergesi olarak yeniden yapılandırılmış, böylece madde sayısı yeniden 18’e ulaşmıştır. Ancak pilot uygulama ve puanlayıcı analizleri sonucunda, gözlemlenmesi pratikte mümkün olmayan 4 madde çıkarılmış ve nihai rubrik 14 maddeden oluşmuştur.

Bu değişikliklerin ayrıntılı gerekçeleri ve uygulama süreci, yöntem bölümünde detaylı biçimde sunulmuştur. Yapılan bu revizyonlar, rubriğin hem geçerliğini hem de uygulamadaki işlevselliğini artırması hedeflenmektedir.

3.4. Tanımlayıcı İstatistikler

Bu bölümde, geliştirilen rubriğin uygulaması sonucunda iki bağımsız puanlayıcı tarafından verilen puanlara ilişkin tanımlayıcı istatistikler sunulmaktadır. Her bir madde için ortalama ($\bar{X}$), standart sapma (SD), minimum ve maksimum değerler Tablo 6'da özetlenmiştir. Bu bulgular, puanlayıcıların değerlendirmeleri ışığında katılımcıların performans düzeylerinin dağılımını ve maddeler arasındaki genel eğilimi ortaya koymaktadır.

Tablo 6. Maddelere ilişkin tanımlayıcı istatistikler.

Madde	Ortalama ($\bar{X}$)	SD	Minimum	Maksimum
Madde1	3.18	0.53	2	4
Madde2	3.27	0.46	2	4
Madde3	3.20	0.42	2	4
Madde5	3.18	0.60	1	4
Madde7	2.86	0.66	1	4
Madde8	3.16	0.70	1	4
Madde9	3.09	0.49	2	4
Madde10	3.23	0.62	2	4
Madde11	2.91	0.82	1	4
Madde12	3.25	0.44	2	4
Madde15	3.18	0.61	2	4
Madde16	3.00	0.77	1	4
Madde17	2.91	0.67	1	4
Madde18	2.91	0.72	1	4

Elde edilen bulgular, tüm maddelerin ortalamalarının 2.86 ile 3.27 arasında değiştiğini göstermektedir. En yüksek ortalama Madde2'de ($\bar{X} = 3.27$), en düşük ortalama ise Madde7'de ($\bar{X} = 2.86$) görülmüştür. Bazı maddelerde (özellikle Madde7, Madde11, Madde17 ve Madde18) görece daha düşük bir başarı düzeyinin ortaya çıktığını göstermektedir. Standart sapma değerlerinin 0.42 ile 0.82 arasında değiştiği ve minimum-maksimum puanların 1–4 aralığında olduğu gözlenmektedir. Özellikle Madde11 (SD = 0.82), Madde16 (SD = 0.77), Madde18 (SD = 0.72) gibi maddelerde varyans nispeten yüksek iken Madde3 (SD = 0.42) ve Madde12 (SD = 0.44) gibi maddelerde standart sapma düşüktür. Bu veriler, iki bağımsız puanlayıcının değerlendirmelerine göre katılımcıların genel olarak rubrik maddelerine yüksek puanlar verdiğini, ancak bazı maddelerde ortalamanın görece daha düşük olduğunu ortaya koymaktadır. Sonuç olarak, geliştirilen rubriğin farklı boyutlarının puanlayıcılar tarafından benzer eğilimlerle değerlendirildiği ve maddeler arası varyansın düşük ila orta düzeyde olduğu söylenebilir.

3. 5. Yapı Geçerliği

Rubriğin yapı geçerliğini belirlemek amacıyla, açıklayıcı faktör analizi (AFA) ön koşulları değerlendirilmiştir. Bu kapsamda, öncelikle örneklem yeterliliğini test etmek için Kaiser-Meyer-Olkin (KMO) testi ve maddeler arası ilişkilerin çok değişkenli analiz için uygun olup olmadığını belirlemek amacıyla Bartlett's Sphericity testi uygulanmıştır. Ayrıca, madde arası korelasyonlar ve çoklu doğrusal bağlantı (multicollinearity) düzeyi incelenmiştir. Analiz sonuçları, ölçeğin faktör analizine uygunluğunu ve maddeler arasındaki ilişkilerin yapısını ortaya koymaktadır. İlgili bulgular Tablo 7'de sunulmuştur.

Tablo 7. Faktör analizi uygunluk testi sonuçları.

Test	Sonuç	Yorum
KMO (Kaiser-Meyer-Olkin)	0.732	Uygun (≥ 0.70)
Bartlett's Test	NaN	Analiz gerçekleştirilemedi; matris tekil
Korelasyonu Yüksek Maddeler	Var	Bazı maddeler arası korelasyon .90'ın üzerinde
Çoklu doğrusal bağlantı	Yok	Varyanslar kabul edilebilir düzeyde

Ölçeğin faktör analizine uygunluğunu değerlendirmek amacıyla yapılan ön analizlerde, Kaiser-Meyer-Olkin (KMO) örneklem yeterlilik katsayısı 0.732 olarak bulunmuştur. Bu değer, ölçme aracının faktör analizine elverişli olduğunu göstermektedir (Kaiser, 1974). Ancak, Bartlett's Sphericity Testi sonucunda korelasyon matrisinin tekil (singular) olduğu belirlenmiş ve bu nedenle faktör analizi matematiksel olarak gerçekleştirilememiştir.

3.6. İç Tutarlılık (Cronbach's Alpha)

Geliştirilen rubriğin puanlama maddeleri arasındaki tutarlılığı değerlendirmek amacıyla Cronbach's Alpha iç tutarlılık katsayısı hesaplanmıştır. Bu analiz, ölçeğin güvenilirliğinin ve maddelerin ölçülen yapıya uygunluğunun belirlenmesi açısından önemlidir. Geliştirilen performans değerlendirme rubriğinin iç tutarlılığını değerlendirmek amacıyla hesaplanan Cronbach's Alpha katsayısı 0.968 olarak bulunmuştur. Bu değer, ölçeğin maddeleri arasında son derece yüksek bir tutarlılığa işaret etmektedir. Genel olarak, 0.70 ve üzerindeki Alpha değerleri kabul edilebilir düzeyde güvenilirlik olarak değerlendirilirken, 0.90'ın üzerindeki değerler ise ölçeğin mükemmel bir iç tutarlılığa sahip olduğunu göstermektedir (Field, 2018). Elde edilen sonuçlar, geliştirilen rubrik maddelerinin aynı yapı altında bütüncül bir şekilde işlediğini ve puanlama sürecinde tutarlı sonuçlar verdiğini göstermektedir.

Her bir madde silindiğinde hesaplanan Cronbach's Alpha değerleri incelendiğinde, ölçekten herhangi bir maddenin çıkarılmasının genel iç tutarlılık katsayısını anlamlı biçimde değiştirmedeği görülmüştür ($\alpha=0.963-0.967$). Bu bulgu, rubrik maddelerinin tamamının ölçeğe bütüncül ve tutarlı bir şekilde katkı sağladığını göstermektedir.

3.7. Cohen's Kappa Analizi

Tablo 8. Rubrik maddelerine ilişkin cohen's kappa analiz sonuçları.

Madde	Cohen's Kappa	Z değeri	p değeri	Yorum
Madde1	0.522	2.40	$p < .001$	Orta
Madde2	1.000	4.64	$p < .001$	Çok iyi
Madde3	0.855	3.99	$p < .001$	Çok iyi
Madde5	0.718	3.47	$p < .001$	İyi
Madde7	0.371	2.22	$p < .001$	Düşük-Orta
Madde8	0.841	3.86	$p < .001$	Çok iyi
Madde9	0.548	2.53	$p < .001$	Orta
Madde10	0.662	2.53	$p < .001$	İyi
Madde11	1.000	4.15	$p < .001$	Çok iyi
Madde12	0.662	2.53	$p < .001$	İyi
Madde15	0.441	1.97	$p < .001$	Orta
Madde16	0.405	1.96	$p < .001$	Orta
Madde17	0.560	2.56	$p < .001$	Orta
Madde18	0.371	2.19	$p < .001$	Düşük-Orta

Tablodaki Cohen's Kappa katsayıları, iki bağımsız puanlayıcının rubrik maddelerine ilişkin değerlendirmelerinde sağlanan uyum düzeyini göstermektedir. Analiz sonuçlarına göre, maddelerin çoğunluğunda Cohen's Kappa değeri .40 ile 1.00 arasında değişmektedir. Özellikle Madde 2, 3, 8 ve 11'de kappa değerinin .80'in üzerinde olması, bu maddelerde puanlayıcılar arası çok yüksek düzeyde tutarlılık elde edildiğini göstermektedir (Landis & Koch, 1977). Madde 5, 10 ve 12 gibi maddelerde ise "iyi" düzeyde (0.61–0.80 arası) uyum sağlanmıştır. Madde 1, 9, 15, 16 ve 17'de uyum "orta" düzeyde (.41–.60 arası) bulunmuştur. Madde 7 ve 18'de ise "düşük-orta" düzeyde (.21–.40 arası) uyum tespit edilmiştir. Tüm maddelerde elde edilen p değerlerinin .05'in altında olması, puanlayıcılar arası uyumun istatistiksel olarak anlamlı olduğunu göstermektedir. Genel olarak bulgular, geliştirilen performans değerlendirme rubriğinin puanlayıcılar arası tutarlılık açısından kabul edilebilir psikometrik özelliklere sahip olduğunu ortaya koymaktadır.

3.8. Sınıf İçi Korelasyon Katsayısı (ICC)

Puanlayıcılar arası tutarlılığın daha bütüncül ve duyarlı biçimde değerlendirilmesi amacıyla, geliştirilen performans değerlendirme rubriği için Sınıf İçi Korelasyon Katsayısı (ICC) analizleri gerçekleştirilmiştir. ICC, iki veya daha fazla puanlayıcı tarafından yapılan bağımsız puanlamaların benzerlik düzeyini ölçmekte olup, özellikle sürekli ölçeklerde puanlayıcılar arası güvenilirliğin belirlenmesinde yaygın olarak kullanılmaktadır (Hallgren, 2012). Elde edilen ICC değerleri ve yorumları aşağıda sunulmuştur.

Tablo 9. İki puanlayıcı arasında ICC sonuçları

ICC Tipi	Açıklama	ICC	F	df1	df2	p-değeri	%95 Güven Aralığı
ICC1	Single rater absolute agreement	0.924	25,29	10	11	p < .001	[0.76, 0.98]
ICC2	Single random rater	0.924	26,37	10	10	p < .001	[0.76, 0.98]
ICC3	Single fixed rater	0.927	26,37	10	10	p < .001	[0.75, 0.98]
ICC1k	Average raters absolute	0.960	25,29	10	11	p < .001	[0.86, 0.99]
ICC2k	Average random raters	0.961	26,37	10	10	p < .001	[0.86, 0.99]
ICC3k	Average fixed raters	0.962	26,37	10	10	p < .001	[0.86, 0.99]

Tablo 9'da sunulan ICC analiz sonuçları incelendiğinde, iki bağımsız puanlayıcının rubrik puanlamaları arasında yüksek düzeyde tutarlılık olduğu görülmektedir. Tek puanlayıcıya dayalı mutlak uyum (ICC1: 0.924), rastgele seçilen puanlayıcı için mutlak uyum (ICC2: 0.924) ve sabit puanlayıcı için mutlak uyum (ICC3: 0.927) katsayıları, tümüyle yüksek güvenilirlik düzeyini işaret etmektedir. Ortalama puanlayıcıya ilişkin katsayılar ise ICC1k = 0.960, ICC2k = 0.961 ve ICC3k = 0.962 olarak hesaplanmış, bu değerler %95 güven aralığında 0.86 ile 0.99 arasında değişmiştir.

Tüm ICC değerlerinin istatistiksel olarak anlamlı bulunması (p < 0.001) ve F değerlerinin yüksek çıkması, puanlayıcılar arası değerlendirmenin hem tekil hem de ortalama bazda oldukça güçlü ve güvenilir olduğunu göstermektedir.

TARTIŞMA, SONUÇ ve ÖNERİLER

Bu araştırmanın temel amacı, fen bilimleri öğretmenleri ve öğretmen adaylarının sınıf içi performanslarını nesnel, geçerli ve güvenilir biçimde değerlendirmek üzere kapsamlı bir performans değerlendirme rubriği geliştirmektir. Bu amaç doğrultusunda, rubrik geliştirme sürecinde kapsamlı doküman analizi, uzman görüşleriyle kapsam geçerliği sağlanması, pilot uygulama ve video gözlemleriyle madde revizyonu, ardından psikometrik analizlerle geçerlik ve güvenilirlik testleri gerçekleştirilmiştir. Bu çalışmada geliştirilen performans değerlendirme rubriğinin, kapsam geçerliği, yapı geçerliği ve güvenilirlik bakımından yüksek psikometrik özellikler sergilediği belirlenmiştir. Elde edilen bulgular, hem ulusal hem de uluslararası literatürde rubrik geliştirme sürecine ilişkin kabul gören temel ilkelerle büyük ölçüde örtüşmektedir. Bu durum, geliştirilen rubriğin alan yazındaki mevcut standartlarla uyumlu ve güvenilir bir değerlendirme aracı olarak kullanılabileceğini göstermektedir. Böylece, hem öğretmen eğitimi hem de öğretmen performansının izlenmesi süreçlerinde yeni bir referans ölçme aracı kullanılabilir.

Kapsam geçerliği analizinde, rubriğin maddelerinin büyük çoğunluğunun uzmanlar tarafından yüksek düzeyde “gerekli” ve “önemli” olarak değerlendirilmesi ($KGO \geq .69$), öğretmen performansının ölçülmesinde bu maddelerin vazgeçilmez olduğunu ortaya koymuştur (bkz. Tablo 5). Özellikle “Planlama”, “Öğrenme Ortamının Planlanması”, “Materyal ve Kaynaklar” ve “Bilimsel Süreç Becerileri” gibi alanlarda tam uzlaşa sağlanması, ölçme aracının teorik altyapısının güçlü olduğuna işaret etmektedir. Bu sonuç, Yeşildağ-Hasancebi ve ark. (2022) tarafından bildirilen yüksek KGO ve KGİ değerleriyle tutarlı olduğu gibi, Permana ve Widodo (2022), Santiani ve ark. (2023) ve Amelia ve ark. (2024) gibi uluslararası çalışmalarla da paralellik göstermektedir. Bu çalışmalar, rubriklerin uzman görüşüne dayalı kapsam geçerliğinin yalnızca ulusal bağlamda değil, farklı eğitim sistemlerinde de kritik öneme sahip olduğunu göstermiştir.

Rubrik geliştirme sürecinde pilot uygulamanın ve nitel uzman dönütlerinin önemi bir kez daha ortaya çıkmıştır. Çalışmamızda, iki bağımsız puanlayıcı tarafından yapılan pilot gözlemler sonucunda, 4 madde ölçekten çıkarılarak rubriğin nihai halinde 14 madde kalmıştır. Disher (2018) da pilot uygulama sürecinde gözlemlenebilirliği düşük maddelerin elenmesinin, ölçeğin hem geçerlik hem de işlevsellik boyutunu güçlendirdiğini raporlamıştır. Bu bulgu, ölçme aracı geliştirme literatüründe sıkça vurgulanan “pratikte uygulanabilirlik” ve “gözlenebilirlik” kriterlerinin, psikometrik kalite kadar önemli olduğunu göstermektedir. Ayrıca, uzmanlardan alınan açık uçlu nitel dönütler doğrultusunda, “Eleştirel düşünme ve problem çözme becerisi” maddesinin iki bağımsız performans göstergesi olarak yeniden yapılandırılması, ölçeğin açıklık ve ayırt edicilik düzeyini artırmıştır. Böylece, kapsam geçerliği değerlendirmesi yalnızca nicel analizlerle sınırlı kalmamış, nitel uzman görüşleriyle de zenginleştirilmiştir.

Yapı geçerliği analizinde, KMO değerinin .732 olarak bulunması ve madde-madde korelasyonlarının büyük ölçüde kabul edilebilir düzeyde olması, rubriğin psikometrik açıdan sağlam bir yapıya sahip olduğunu göstermektedir. Bartels ve ark. (2019), Alias ve Osman (2015) ve Amelia ve ark. (2024) gibi çalışmalarda da benzer şekilde yapı geçerliği, faktör analizi ile desteklenmiş ve ölçme araçlarının çok boyutlu yapısının uygunluğu vurgulanmıştır. Çalışmamızda Bartlett’s Sphericity testinin teknik nedenlerle uygulanamaması, matrisin tekil olmasından kaynaklanmakla birlikte, diğer istatistiksel göstergeler rubriğin faktör analizine uygun olduğunu göstermiştir. Bu bulgu, Santiani ve ark. (2023) gibi çalışmalarda raporlanan çoklu madde ilişkileriyle de örtüşmektedir.

Geliştirilen rubriğin güvenilirliği, hem iç tutarlılık (Cronbach’s Alpha = .968) hem de puanlayıcılar arası güvenilirlik (Cohen’s Kappa ve ICC) ile yüksek düzeyde saptanmıştır. Özellikle, Cronbach’s Alpha değerinin .968 gibi olağanüstü bir düzeye ulaşması, ölçeğin tutarlı ve güvenilir biçimde kullanılabileceğini göstermektedir. Aktaş ve Özmen (2022) (.942), Huda ve ark. (2020) (.900) ve Bartels ve ark. (2019) (.56–.79) gibi çalışmalarda raporlanan Alpha değerleriyle kıyaslandığında, bu çalışma ulusal ve uluslararası standartların üzerinde bir iç tutarlılık sunmaktadır. Puanlayıcılar arası tutarlılığa bakıldığında ise, Cohen’s Kappa katsayılarının orta ve üzerinde, ICC değerlerinin ise .92–.96 arasında değişmesi, rubriğin farklı puanlayıcılar tarafından güvenle kullanılabileceğini göstermektedir. Permana ve Widodo (2022) (.96) ve Aumann ve ark. (2023) (.873) gibi literatürdeki yüksek ICC bulguları ile uyumlu sonuçlar elde edilmiştir.

Çalışmanın genel katkısı, Millî Eğitim Bakanlığı’nın öğretmen yeterlikleriyle ve güncel eğitim uygulamalarıyla uyumlu, nesnel ve güvenilir bir performans değerlendirme aracı sunmasıdır. Rubriğin kapsam ve yapı geçerliğinin titizlikle sağlanması, yüksek iç tutarlılığı ve puanlayıcılar arası güvenilirliği ile mevcut alanyazına anlamlı bir katkı sunduğu söylenebilir. Bununla birlikte, örneklem büyüklüğü ve uygulamanın yalnızca belirli bir coğrafi bölgede gerçekleştirilmiş olması, bulguların genellenebilirliği açısından bir sınırlılık olarak değerlendirilmelidir. Gelecekte, daha geniş ve farklı bölgelerde, farklı branşlarda ve daha fazla puanlayıcı ile benzer çalışmalar yürütülerek bulguların desteklenmesi önerilmektedir.

Sonuç olarak, bu çalışma öğretmen performansının çok boyutlu, nesnel ve güvenilir biçimde değerlendirilmesine olanak tanıyan, ulusal ve uluslararası standartlarla uyumlu bir rubrik geliştirilmesinde önemli bir adım niteliği taşımaktadır. Elde edilen psikometrik bulgular, hem

literatürdeki güncel yaklaşımlarla uyumlu hem de Türk eğitim sisteminin özgün gereksinimlerine yanıt verecek niteliktedir.

Bu çalışmada geliştirilen performans değerlendirme rubriği, fen bilimleri öğretmenleri ve öğretmen adaylarının mesleki yeterliklerini daha nesnel ve çok boyutlu biçimde izleme ve geliştirme imkânı sunmaktadır. Ölçeğin öğretmenler tarafından öz değerlendirme veya meslektaş geri bildirimi amacıyla kullanılması, bireysel gelişim ve sürekli mesleki öğrenmeye katkı sağlayabilir. Eğitim yöneticileri ve okul idarecileri, rubriği öğretmen performansını izleme, geribildirim verme ve hizmet içi eğitim planlamalarında objektif bir araç olarak kullanabilirler. Üniversitelerdeki öğretmen yetiştirme programlarında ise, öğretmenlik uygulamaları dersi kapsamında öğretmen adaylarının gelişimi daha sistematik ve şeffaf biçimde takip edilebilir. Bununla birlikte, geliştirilen rubriğin farklı örneklem gruplarıyla uygulanarak psikometrik özelliklerinin (geçerlik ve güvenilirlik) çeşitli bağlamlarda tekrar test edilmesi, ölçeğin alana katkısını daha da güçlendirecek ve genellenebilirliğini artıracaktır.

Etik ve Çıkar Çatışması

Bu çalışma, Dokuz Eylül Üniversitesi Eğitim Bilimleri Enstitüsü Matematik ve Fen Bilimleri Eğitimi Anabilim Dalı Fen Bilgisi Öğretmenliği Doktora Programı kapsamında yürütülen “Fen Bilimleri Öğretmen ve Öğretmen Adaylarına Yönelik Bir Performans Değerlendirme Rubriği Geliştirilmesi: Geçerlik ve Güvenirlik Çalışması” başlıklı doktora tezinden üretilmiştir. Araştırma için Dokuz Eylül Üniversitesi Etik Kurulu’ndan 12/12/2023 tarihli ve 4 sayılı kararla etik onay alınmış, ayrıca Millî Eğitim Bakanlığı ve İzmir İl Millî Eğitim Müdürlüğü’nden 25/03/2025 tarihli ve 99466338 sayılı izinlerle gerekli yasal izinler temin edilmiştir. Çalışma süreci boyunca bilimsel araştırma ve yayın etiği ilkelerine tam olarak uyulmuştur. Yazarlar arasında herhangi bir çıkar çatışması bulunmamaktadır.

KAYNAKÇA

- Aktaş, M., & Özmen, H. (2022). Fen bilgisi öğretmen adaylarına yönelik rubrik geliştirme: Geçerlik ve güvenilirlik çalışması. *Journal of Turkish Science Education*, 19(2), 456-473.
- Alias, N., & Osman, R. (2015). Learning outcomes and reliability of rubric scoring for online discussion: Peer-and self-assessment. *Procedia-Social and Behavioral Sciences*, 204, 309-316.
- Amelia, R., Santiana, S., Permana, I., & Syawaludin, A. (2024). The development of a rubric-based assessment instrument for science teacher performance. *Journal of Science Teacher Education*, 35(1), 102-118.
- Antony, J. S., & Paidi, S. (2019). Development and validation of a performance assessment rubric for science teachers. *International Journal of Science Education*, 41(5), 605-627.
- Aumann, M., Schnebel, S., & Weitzel, T. (2023). Development and validation of the EnTPACK rubric: Assessing technological pedagogical content knowledge of science teachers. *Teaching and Teacher Education*, 125, 104044.
- Bartels, J., Schmiedebach, M., & Ploetzner, R. (2019). Development and validation of a performance-based rubric for assessing science teachers' skills in Germany and Australia. *International Journal of Science Education*, 41(7), 927-947.
- Brookhart, S. M. (2013). *How to create and use rubrics for formative assessment and grading*. ASCD.
- Chasteen, S. V., & Scherr, R. E. (2020). Teacher performance assessment using video observation. *Physical Review Physics Education Research*, 16(2), 020123.
- Creswell, J. W. (2017). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). Sage.
- Darling-Hammond, L. (2000). Teacher quality and student achievement: A review of state policy evidence. *Education Policy Analysis Archives*, 8, 1-44.
- DeVellis, R. F. (2017). *Scale development: Theory and applications* (4th ed.). Sage.
- Disher, C. (2018). *The development and validation of a performance evaluation instrument for speech-language pathologists*. Doctoral dissertation, University of California, Berkeley.
- Edwards, A. (2017). Assessing teacher performance: A practical guide. *Educational Assessment*, 22(1), 15-32.
- Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2012). *How to design and evaluate research in education* (8th ed.). McGraw-Hill.
- Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, 8(1), 23-34. <https://doi.org/10.20982/tqmp.08.1.p023>

- Huda, N., Sari, D. P., & Yuliana, M. (2020). Developing a rubric for laboratory practice: Validity and reliability studies. *Journal of Science Education*, 21(3), 233–245.
- Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, 2(2), 130–144. <https://doi.org/10.1016/j.edurev.2007.05.002>
- Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika*, 39, 31–36.
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159–174.
- Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology*, 28(4), 563–575.
- Marzano, R. J. (2012). *The art and science of teaching: A comprehensive framework for effective instruction*. ASCD.
- MEB (2008). *Milli Eğitim Bakanlığı Fen ve Teknoloji Öğretmeni Özel Alan Yeterlikleri*, 25 Temmuz 2008 tarih ve 2391 sayılı onayı, <http://otmg.meb.gov.tr/alanfen.html>.
- MEB (2017). *Öğretmenlik mesleği genel yeterlikleri*. Ankara: Öğretmen Yetiştirme ve Geliştirme Genel Müdürlüğü.
- MEB. (2023). *Öğretmenlik meslek kanunu tasarısı*. Millî Eğitim Bakanlığı, Ankara.
- NBPTS. (2022). *National Board for Professional Teaching Standards: Science standards*.
- Ofsted. (2019). *Education inspection framework. Office for Standards in Education*, United Kingdom.
- PACT. (2018). *Performance Assessment for California Teachers*. Stanford Center for Assessment, Learning and Equity.
- Permana, I., & Widodo, A. (2022). Content validity and reliability of a science teacher performance rubric in Indonesia. *Indonesian Journal of Science Education*, 11(4), 482–491.
- Petrie, K. (2023). 21st century skills in science education. *Journal of Curriculum Studies*, 55(1), 100–120.
- Santiani, S., Amelia, R., Permana, I., & Syawaludin, A. (2023). Development of an interdisciplinary thinking skills rubric (IITSSL). *International Journal of Instruction*, 16(2), 243–260.
- Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher*, 15(2), 4–14.
- Stevens, D. D., & Levi, A. J. (2013). *Introduction to rubrics: An assessment tool to save grading time, convey effective feedback, and promote student learning* (2nd ed.). Stylus.
- Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research, and Evaluation*, 9(4), 1–19.
- Syawaludin, A., Santiani, S., Amelia, R., & Permana, I. (2024). Science teacher performance rubric development: Psychometric evaluation. *Journal of Science Teacher Education*, 35(2), 130–147.
- Yeşilbaş-Hasancebi, F., Demirci, C., & Yılmaz, M. (2022). Fen bilimleri öğretmen adayları için performans değerlendirme rubriği: Geliştirilmesi, geçerliği ve güvenilirliği. *Turkish Journal of Science Education*, 19(4), 879–896.
- Yurdugül, H. (2005). Ölçme aracı geliştirme çalışmalarında kapsam geçerliği için kapsam geçerlik indeksi kullanımı. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, 1(1), 25–28.
- YÖK. (2021). *Araştırma ve yayın etiği yönergesi*. Yükseköğretim Kurulu. <https://www.yok.gov.tr/Sayfalar/Hakkinda/yayin-etik.aspx>

EXTENDED ABSTRACT

This study aims to develop a valid and reliable performance assessment rubric to evaluate the professional competencies of science teachers and pre-service teachers in alignment with both national and international teaching standards. Ensuring quality assurance in education requires objective and multidimensional tools capable of measuring teachers' pedagogical performance comprehensively. Teacher performance evaluation should extend beyond subject-matter knowledge to include pedagogical practice, communication, technology integration, material use, and interdisciplinary thinking. Accordingly, this study addresses a critical gap in the literature by presenting a psychometrically sound and practically applicable rubric for assessing science teaching performance. The research employed a methodological design based on the descriptive survey model and followed seven systematic stages: (1) document analysis, (2) development of a draft rubric, (3) linguistic expert review, (4) content validity testing, (5) item revision, (6) pilot implementation through video-based classroom observations, and (7) final evaluation. Initially, the “General and Special Field Competencies for the Teaching Profession” published by the Turkish Ministry of National Education (2008; 2017),

along with international standards such as AITSL, NBPTS, Ofsted, and PACT, were analyzed to create a comprehensive item pool focusing on observable teaching behaviors. Two Turkish language experts reviewed the draft to ensure linguistic clarity and content appropriateness. Content validity was examined using the Lawshe (1975) technique, based on expert ratings from 13 participants (3 academicians and 10 in-service science teachers). Items were evaluated on a three-point scale (“not essential,” “useful but not essential,” “essential”). The Content Validity Ratios (CVR) ranged from 0.54 to 1.00, and the overall Content Validity Index (CVI) was 0.83, indicating strong content representativeness. Following expert feedback, the “Kemalism” item was removed, while “critical thinking and problem solving” was divided into two separate items to enhance precision. A pilot study was conducted with 15 science teachers and 11 pre-service teachers, whose classroom sessions were video recorded and evaluated by two independent raters. Four items—“curiosity toward environmental exploration,” “nature and history of science,” “safety precautions,” and “special needs adaptation”—were removed due to low observability in classroom settings. The final version included 14 items, ensuring practical applicability. Data analyses conducted via SPSS 25 revealed that the Kaiser-Meyer-Olkin (KMO) value was 0.732, supporting the adequacy of the dataset for factor analysis. Although Bartlett’s Test of Sphericity could not be performed due to matrix singularity, the inter-item correlations were satisfactory, confirming the rubric’s construct validity. The internal consistency coefficient (Cronbach’s Alpha) was found to be exceptionally high ($\alpha = .968$), indicating strong coherence among items and excellent reliability. Inter-rater reliability was determined using Cohen’s Kappa and Intraclass Correlation Coefficient (ICC) analyses. Kappa values ranged from .37 to 1.00, with several items (2, 3, 8, and 11) exceeding .80, showing “very good” agreement. The ICC values ranged from .92 to .96 ($p < .001$), confirming a high level of consistency between raters. Overall, the findings demonstrate that the developed rubric exhibits strong psychometric properties in terms of validity, reliability, and feasibility. The rubric enables the multidimensional evaluation of teachers’ professional competencies across domains such as lesson planning, instructional process, material utilization, scientific process skills, communication, attention to individual differences, and technology integration. The use of behaviorally anchored, observable indicators enhances objectivity and fairness in teacher performance assessment. In the discussion, the results were compared with relevant studies (Aktaş & Özmen, 2022; Aumann et al., 2023; Permana & Widodo, 2022), revealing that the reliability coefficients obtained in this study ($\alpha = .968$; ICC = .92–.96) surpass those reported in similar research. The removal of items with limited observability strengthened the rubric’s practical usability and psychometric robustness. The study thus contributes a comprehensive, standardized, and evidence-based assessment tool that aligns with both the Ministry of National Education’s teacher competency framework and international quality standards. The rubric can be effectively used for monitoring teachers’ professional growth, facilitating formative and summative assessments, providing feedback, and structuring professional development programs. Furthermore, it offers a valid and objective framework for assessing pre-service teachers during their practicum experiences. Despite its strong psychometric evidence, the study is limited by its relatively small and region-specific sample, which may affect generalizability. Future research should replicate and expand the application of the rubric across diverse regions and larger samples to further validate its structure. In conclusion, this study presents a rigorously developed and empirically validated rubric that provides an objective, multidimensional, and reliable framework for evaluating science teachers’ classroom performance. The instrument contributes to enhancing quality assurance in teacher education and supports the development of an evidence-based evaluation culture grounded in transparency and professional growth.