

ARAŞTIRMA

Açık Erişim

**Türkçe Öğretmeni Adaylarının Sınav Hazırlama Becerilerine Yönelik
Rubrik Geliştirme Çalışması**

Efecan Karagöl*

Öz

Bu araştırma, Türkçe öğretmeni adaylarının sınav hazırlama ve değerlendirme becerilerini nesnel, sistematik ve ölçülebilir biçimde ortaya koymak amacıyla analitik bir rubrik geliştirmeyi amaçlamaktadır. Öğretmen adaylarının ölçme-değerlendirme bilgi ve becerilerinin hem öğretim sürecinin niteliğini hem de öğrenci başarısını doğrudan etkilediği dikkate alındığında, sınav tasarımı sürecini bütüncül olarak değerlendirebilecek bir ölçme aracına ihtiyaç duyulmaktadır. Rubrik geliştirme süreci, alan yazın incelemesi, ölçütlerin belirlenmesi, taslak formun oluşturulması ve uzman görüşleri doğrultusunda kapsam geçerliğinin sağlanması aşamalarını içermektedir. Lawshe analizi sonucunda tüm maddelerin kapsam geçerlik ölçütünü karşıladığı belirlenmiştir. Rubriğe ilişkin puanlayıcılar arası ilişki Spearman korelasyon katsayısı ile incelenmiş ve katsayıların ,708 ile ,965 arasında değiştiği görülmüştür. Güvenirlik analizleri kapsamında Cronbach's Alpha katsayısı ,772 olarak hesaplanmış ve rubriğin yeterli düzeyde iç tutarlılık gösterdiği belirlenmiştir. Ayrıca Cohen's Kappa katsayılarının tüm maddelerde iyi düzeyde uyuma işaret ettiği görülmüştür. Elde edilen bulgular, geliştirilen rubriğin Türkçe öğretmeni adaylarının sınav hazırlama ve değerlendirme becerilerini değerlendirmede kullanılabilir bir ölçme aracı olduğunu göstermektedir.

Anahtar Kelimeler: Sınav hazırlama, rubrik geliştirme, Türkçe öğretmeni adayları, analitik rubrik, geçerlik, güvenilirlik.

Makale Geçmişi

Geliş: 13/02/2026

Düzeltilme: 19/03/2026

Kabul: 09/05/2026

* * Prof. Dr., Zonguldak Bülent Ecevit Üniversitesi, 0000-0003-0331-8009, Zonguldak, Türkiye, efecankaragol@gmail.com

Atıf için: Karagöl, E. (2026). Türkçe öğretmeni adaylarının sınav hazırlama becerilerine yönelik rubrik geliştirme çalışması. *OJCES*, 4(1), 87-108. <https://doi.org/10.5281/zenodo.20802678>

Giriş

Türkçe öğretmeni adaylarının sınav hazırlama sürecinde sahip olmaları gereken beceriler, nitelikli eğitim ortamlarının oluşturulması açısından önemli bir yere sahiptir. Söz konusu becerilerin kuramsal bilginin ötesine geçerek ölçülebilir ve geliştirilebilir uygulamalarla desteklenmesi önem taşır. Bu çerçevede öğretmen adaylarının ölçme-değerlendirme yeterliklerini ortaya koyabilecek, düzenli ve güvenilir bir değerlendirme aracına ihtiyaç duyulmaktadır. Rubrikler ise bireylerin performanslarını belirli ölçütler doğrultusunda değerlendirme imkânı sunarak bu ihtiyaca yanıt verebilecek etkili araçlar arasında yer almaktadır.

Bir öğretmenin etkili bir öğretim süreci yürütebilmesi, öğrencilerin düşünmesini, anlamasını ve öğrendiklerini transfer etmesini destekleyen nitelikli sorular oluşturabilmesiyle doğrudan bağlantılıdır (Yüceer, 2024). Bu bağlamda öğretmenin değerlendirme süreçlerini amacına uygun biçimde yürütebilmesi, bu süreçleri öğrenme alanının hedefleriyle uyumlu olacak şekilde planlamasını, yapılandırmasını ve uygulamasını gerektirir. Sınıf içinde kullanılan sorular yalnızca öğrencilerin mevcut bilgi düzeylerini ortaya koymakla kalmaz aynı zamanda onların anlamı inşa etmelerine ve öğrenme sürecini derinleştirmelerine katkı sağlar. Bu işlevsel yönüyle sorular, öğrencilerin öğretileni ne ölçüde hatırladığını, kavradığını ve uygulayabildiğini belirlemeye hizmet eden iki temel işlevi yerine getirir (Kintsch, 2005). Bununla birlikte etkili sorular öğrencilerin üst düzey düşünme becerilerini harekete geçirerek eleştirel ve yaratıcı düşüncelerine de zemin hazırlar. Ayrıca öğretmenin sınıf içi etkileşimi yönlendirmesine, öğrenci ihtiyaçlarını belirlemesine ve öğretim sürecini dinamik biçimde uyarlamasına olanak tanır. Bu noktada öğretmenlerin ölçme ve değerlendirme okuryazarlığına sahip olmaları, hem soruların niteliğini artırmaları hem de ölçme-değerlendirme süreçlerini daha bilinçli ve etkili biçimde yürütebilmeleri açısından kritik bir gereklilik hâline gelmektedir.

Ölçme ve değerlendirme okuryazarlığı, öğretmenlerin mesleki yeterliklerinin temel bileşenlerinden biri olarak kabul edilmesine rağmen bu alanın öğretmen yetiştirme programlarında çoğu zaman geri planda kaldığı görülmektedir (Khan, 2023, s. 345). Oysa ölçme ve değerlendirme stratejilerini tartışma, uygulama ve tasarlama konusundaki bilgi ve becerileri kapsayan bu okuryazarlık, etkili bir öğretim sürecinin vazgeçilmez unsurudur (Mertler, 2004). Nitekim Stiggins (1995), ölçme-değerlendirme okuryazarı olan öğretmenlerin neyi değerlendirecekleri, öğrencilerin bilgi ve yeteneklerini en iyi hangi yöntemlerle belirleyebilecekleri, ölçme süreçlerini olumsuz etkileyen etmenlerin neler olduğu ve bu etmenlerin nasıl önlenebileceği gibi kritik konularda yeterli donanımına sahip olmaları gerektiğini vurgulamaktadır. Bu yeterliklerin eksikliği, öğretmenlerin kendilerinden beklenen değerlendirme performansını sergilemelerini güçleştirmekte ve sınıf içi uygulamaların etkililiğini azaltmaktadır (Kolcuoğlu ve Boylu, 2024). Ölçme-değerlendirme yeterliğine sahip bir öğretmen ise öğrencilerin ne öğrendiklerini belirlemek için uygun ölçme-değerlendirme etkinlikleri planlayabilir, süreci etkili biçimde yürütebilir ve elde edilen sonuçlara göre öğrencilere gelişimlerini destekleyen nitelikli dönütler sunabilir. Bu bağlamda öğretmenin ölçme ve değerlendirme alanındaki bilgi ve becerilerini pedagojik bilgisi, içerik bilgisi ve sınıfın öğrenme ortamı ile bütünleştirebilmesi hem öğretici hem de ölçmeci rolünü (Karaman, 2022) başarıyla yerine getirmesinin temel koşulu olarak öne çıkmaktadır. Belirtilen bütünleşik yeterlikler, öğretmenin hem öğretici hem de ölçmeci rolünü etkili biçimde yerine getirmesini sağlarken sınav hazırlamada dikkat edilmesi gereken hususların neden titizlikle ele alınması gerektiğini de açıkça göstermektedir.

Türkçe öğretmeni yetiştirme sürecinde ölçme ve değerlendirme bilgi ve becerileri, program yapısının bütününde önemli bir yer tutsa da ders içeriklerine bakıldığında bu alanın çoğu zaman parçalı

ve kuramsal bir çerçevede ele alındığı görülmektedir. Programda yer alan “Eğitimde Ölçme ve Değerlendirme” dersi, ölçme-değerlendirmeye ilgili temel kavramları, ölçme araçlarının psikometrik özelliklerini, başarı testi geliştirme süreçlerini, madde analizi, sonuçların yorumlanması ve not verme gibi teknik boyutları kapsamaktadır (YÖK, 2018). Buna karşın alan eğitimi seçmeli derslerinden “Türkçe Öğretiminde Sınav Hazırlama ve Değerlendirme” dersi; farklı yaş grupları ve dil düzeyleri için sınav türlerini, dil becerilerini ölçmeye yönelik ilkeleri, okuma-yazma-dinleme-konuşma ile kelime bilgisi ve dil bilgisine ilişkin soru türlerini, sınav hazırlama tekniklerini ve değerlendirme ölçütlerini ele almaktadır (YÖK, 2018). Öte yandan her iki dersin içeriği birlikte değerlendirildiğinde öğretmen adaylarına tüm bu bilgileri bütüncül bir sınav tasarım sürecinde bir araya getirecek, kapsam geçerliğini ve madde niteliğini uygulamalı biçimde deneyimletecek sistematik bir yapı sunulmadığı görülmektedir. Bu eksiklik adayların hem psikometrik ilkeleri hem de dil becerilerine özgü soru yazma bilgilerini tutarlı bir sınav hazırlama sürecinde bütünleştirmelerini zorlaştırmakta, sınav tasarlama konusunda belirsizlikler ve uygulama yetersizlikleri ortaya çıkarmaktadır. Dolayısıyla bu durum, öğretmen adaylarının sınav hazırlama performansını nesnel biçimde değerlendirecek ve gelişim alanlarını görünür kılacak nitelikli bir rubriğin geliştirilmesini gerekli hâle getirmektedir.

2024 Ortaokul Türkçe Dersi Öğretim Programı, ölçme ve değerlendirmenin öğretim sürecinin ayrılmaz bir bileşeni olduğunu güçlü biçimde vurgulamaktadır. Programda sınıf içi değerlendirmenin temel amacının öğrenme sürecini desteklemek olduğu belirtilmektedir. Bu doğrultuda ölçme araçlarının açık, anlaşılır, öğrencinin düzeyine uygun ve farklı değerlendirme görev türlerini içerecek biçimde tasarlanması gerektiği ifade edilmektedir. Ayrıca program, öğrencinin sınıf içinde deneyimlemediği görev biçimlerinin sonuç değerlendirmede kullanılmasının başarısızlığa yol açabileceğini, bu nedenle değerlendirme süreçlerinin süreklilik ve tanıdıklık ilkesine dayanması gerektiğini vurgulamaktadır. Okuma, dinleme, konuşma ve yazma becerilerine yönelik ölçme türlerinin ayrıntılı biçimde tanımlandığı programda, üretim görevlerinden sonra kontrol listeleri ve dereceli puanlama anahtarlarıyla değerlendirme yapılması ve öğrenciye sürece odaklanan nitelikli geri bildirim verilmesi gerektiği belirtilmektedir (MEB, 2024). Bu çerçevede, öğretmenin sınav hazırlama sürecinde yalnızca madde yazmakla değil öğrenme çıktılarıyla tutarlı, beceri temelli, süreç odaklı ve açık ölçütlere dayalı bir değerlendirme sistemi kurmakla yükümlü olduğunu göstermektedir. Ancak program bu ilkeleri somut bir sınav hazırlama sürecinde nasıl bütünleştireceğine ilişkin uygulamalı standartlar veya örnek sınav tasarımları sunmadığından, özellikle öğretmen yetiştirme sürecinde sınav hazırlama yeterliklerinin sistematik ve yapılandırılmış bir biçimde geliştirilmesi ve izlenmesi güçleşmektedir. Bu durum, öğretmen adaylarının ölçme-değerlendirme ilkelerini bütüncül bir sınav tasarımına dönüştürürken rehberlik edecek araçlara duyduğu ihtiyacı daha da belirgin hâle getirmektedir.

Her ne kadar öğretmen adaylarının sınav hazırlamaya ve ölçme-değerlendirmeye ilişkin görüşlerini ele alan çalışmalar (Çiçek, 2024; Sallabaş & Kiler, 2024; Yüceer, 2024), ölçme-değerlendirme yeterliklerini inceleyen araştırmalar (Doğan Kahtalı & Çelik, 2019; Karatay & Dilekçi, 2019; Türkben, 2022) ve sınav içeriklerini analiz eden çalışmalar (Ömeroğlu, 2024; Ünlü, Öztürk & Taha, 2014) bulunsun da alan yazında Türkçe öğretmeni adaylarının sınav hazırlama ve değerlendirme becerilerini doğrudan ölçmeye yönelik bir ölçme aracının olmadığı görülmektedir. Söz konusu eksiklik, öğretmen adaylarının sınav tasarlama süreçlerindeki yeterliklerinin nesnel, sistematik ve karşılaştırılabilir biçimde ortaya konmasını güçleştirmektedir. Bu nedenle araştırmanın amacı Türkçe öğretmeni adaylarının sınav hazırlama ve değerlendirme becerilerini geçerli ve güvenilir biçimde ölçebilecek analitik bir rubrik geliştirmektir.

Yöntem

Çalışma Grubu

Rubrik geliştirme sürecinde üç ayrı katılımcı grubuyla çalışılmıştır. Kapsam geçerliğini sağlamak amacıyla Türkiye'nin çeşitli illerinde görev yapan Prof. Dr., Doç. Dr., Dr. Öğr. Üyesi unvanına sahip akademisyenlerden oluşan uzman bir grup sürece dâhil edilmiştir. Çalışmada yer alan katılımcıların özellikleri Tablo 1'de sunulmaktadır.

Tablo 1. *Uzman Grubunun Özellikleri*

		Dr. Öğr. Üyesi	Doç. Dr.	Prof. Dr.
Cinsiyet	Kadın	2	1	-
	Erkek	1	2	1
Toplam		3	3	1

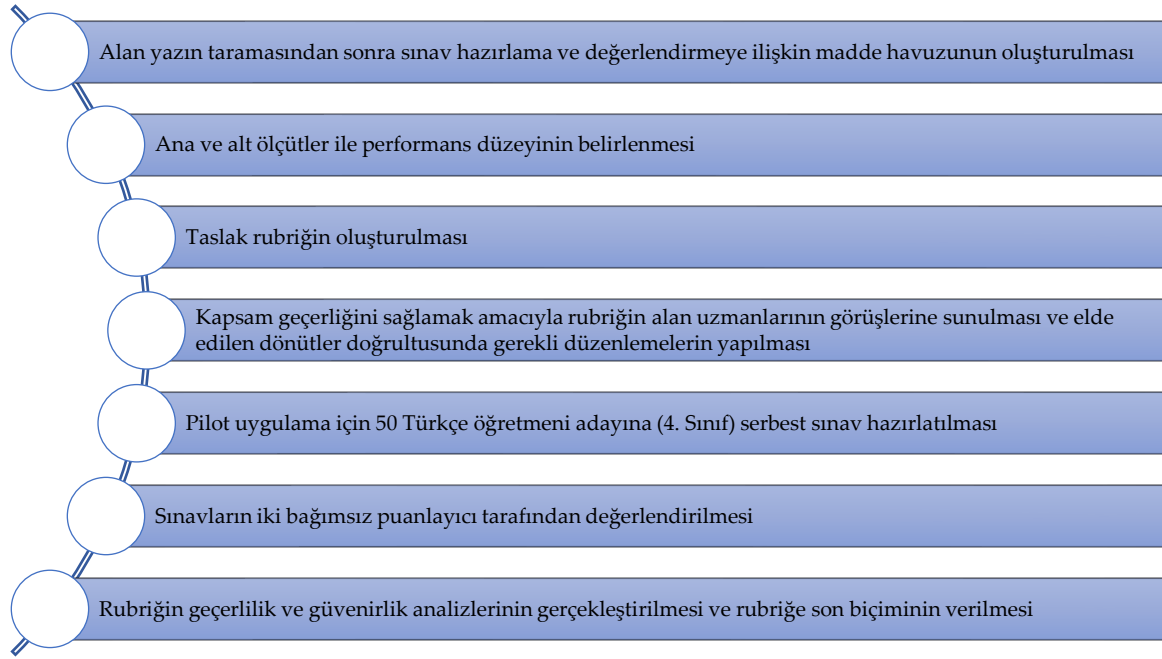
Tablo 1'de görüldüğü gibi araştırmaya Dr. Öğr. Üyesi (f=3), Doç. Dr. (f=3) ve Prof. Dr. (f=1) katılmıştır. Araştırmanın yapı geçerliliği ve güvenilirlik çalışmaları için Batı Karadeniz'deki bir üniversitenin Türkçe eğitimi anabilim dalında 4. sınıfta okuyan 50 öğretmen adayıyla çalışılmıştır. Bu kişiler kolay ulaşılabilir (elverişli) örneklem yöntemiyle seçilmiştir. Ayrıca çalışmanın güvenilirlik boyutunda puanlayıcılar arası uyum için araştırmacı ile 1 uzman tarafından eş zamanlı puanlama yapılmıştır.

Veri Toplama Araçları

Çalışmada veri toplama süreci, taslak rubriğin kapsam geçerliğinin belirlenmesi amacıyla uzman görüşlerine sunulmasıyla başlamıştır. Uzmanlardan elde edilen dönütler doğrultusunda rubrikte gerekli düzenlemeler yapılmış ve araç değerlendirme için hazır hâle getirilmiştir. Bu aşamanın ardından çalışma grubunda yer alan öğretmen adaylarının hazırladığı sınavlar, taslak rubrik kullanılarak iki bağımsız puanlayıcı tarafından puanlanmıştır. Elde edilen veriler üzerinden gerçekleştirilen geçerlik ve güvenilirlik analizleri sonucunda rubrikte son düzenlemeler yapılmış ve böylece kullanılabilir bir nihai değerlendirme aracı oluşturulmuştur.

Verilerin Toplanması ve Analizi

Türkçe öğretmeni adaylarının sınav hazırlama ve değerlendirme becerilerini ölçmek amacıyla geliştirilen rubriğin oluşturulma sürecinde Andrade'nin (1997) önerdiği temel aşamalar dikkate alınmıştır. Bu doğrultuda önce konuya ilişkin alan yazın kapsamlı biçimde incelenmiş, ardından değerlendirilecek ölçütler ile performans düzeyleri belirlenmiştir. Belirlenen unsurlar çerçevesinde taslak rubrik hazırlanmış ve uygulamaya alınmıştır. Uygulama sürecinde elde edilen dönütler ışığında rubrikte gerekli düzenlemeler yapılmış, sonrasında geçerlik ve güvenilirlik analizleri yürütülmüştür. Tüm bu aşamaların tamamlanmasıyla birlikte rubrik nihai biçimine kavuşturularak kullanılabilir bir değerlendirme aracı hâline getirilmiştir. Andrade'nin önerileri doğrultusunda araştırmanın süreci Şekil 1'de gösterilmiştir:



Şekil 1. Rubrik geliştirme süreci.

- Şekil 1’den de izlenebileceği gibi, gerçekleştirilen alan yazın incelemesi sonrasında Türkçe öğretmeni adaylarının sınav hazırlama becerilerini değerlendirmeye yönelik kapsamlı bir madde havuzu oluşturulmuştur.
- Çalışma kapsamında geliştirilen rubrik iki temel boyut üzerine kurulmuştur: “Sınav tasarımı becerisi” ve “değerlendirme ve nesnellik becerisi”. “Sınav tasarımı becerisi” boyutu; beceri alanlarının öğrenme çıktılarına uygunluğu, soru çeşitliliği, soruların bilişsel düzeylere dengeli dağılımı, dil ve anlatım kurallarına uygunluk ile yönergeler alt ölçütlerinden oluşmaktadır. “Değerlendirme ve nesnellik becerisi” boyutu ise değerlendirme ölçütleri ve nesnellik alt ölçütlerini kapsamaktadır. Rubrikte her bir ölçüt, “İyi (4)”, “Orta (3)”, “Geliştirilmeli (2)” ve “Yetersiz (1)” olmak üzere dört düzeyde puanlanmaktadır. Bu yapı doğrultusunda bir sınav ürününden elde edilebilecek toplam puan aralığı 7 ile 28 arasında değişmektedir. Rubrikten elde edilen puanların yorumlanmasında yalnızca toplam puanlar değil, ölçütlere ait ortalama puanlar da dikkate alınmıştır. Her bir ölçütün 1 ile 4 arasında puanlanması nedeniyle değerlendirme sürecinde ortalama puanlar üzerinden yorum yapılmıştır. Ortalama puan aralıklarının belirlenmesinde aralık katsayısı esas alınmış ve performans düzeyleri buna göre sınıflandırılmıştır.
- Tablo 2’de ortalama puanlara göre belirlenen performans düzeyleri sunulmaktadır. Buna göre 1,00–1,75 arası “yetersiz”, 1,76–2,50 arası “geliştirilmeli”, 2,51–3,25 arası “orta” ve 3,26–4,00 arası “iyi” düzeyi ifade etmektedir. Ortalama puan aralıklarının belirlenmesinde aralık katsayısı dikkate alınmıştır. Aralık katsayısı, en yüksek ve en düşük puan arasındaki farkın performans düzeyi sayısına bölünmesiyle hesaplanmıştır $[(4-1)/4 = 0,75]$
- Ortalama puan aralıklarına göre performans düzeyleri şu şekilde belirlenmiştir:
 1,00–1,75 arası yetersiz;
 1,76–2,50 arası geliştirilmeli;
 2,51–3,25 arası orta;
 3,26–4,00 arası iyi.

Bu sınıflandırma, rubrikten elde edilen ortalama puanların yorumlanmasını kolaylaştırmak amacıyla kullanılmıştır.

- Rubriğin kapsam geçerliliğini belirlemek amacıyla Lawshe yöntemi kullanılmış ve uzmanlara “madde hedeflenen yapıyı ölçüyor”, “madde kısmen ilişkili ancak gereksiz” ve “madde ilgili yapıyı ölçmüyor” seçeneklerini içeren bir belirtke tablosu sunulmuştur.
- Uzmanlardan elde edilen geri bildirimler doğrultusunda rubrik maddelerinde gerekli düzenlemeler yapılmış, sonrasında kapsam geçerlik oranı (KGO) ile kapsam geçerlik indeksi (KGİ) hesaplanmıştır.
- Pilot uygulama sürecinde, 4. sınıfta öğrenim gören ve “Türkçe Öğretiminde Sınav Hazırlama ve Değerlendirme” dersini henüz almamış 50 Türkçe öğretmeni adayına sınav hazırlama görevi verilmiştir.
- Bu sınavlar, puanlayıcılar arası tutarlılığı incelemek amacıyla iki bağımsız değerlendirici (P1 ve P2) tarafından puanlanmıştır.
- Rubriğe ilişkin puanlayıcılar arası ilişkiyi incelemek amacıyla Spearman korelasyon katsayısı kullanılmıştır.
- Güvenirlilik analizlerinde iki puanlayıcının verdiği puanlar arasındaki iç tutarlılık Cronbach’s Alpha katsayısı ile belirlenmiştir. Ayrıca puanlayıcılar arasındaki uyumu ortaya koymak için Cohen’in Kappa katsayısı hesaplanmıştır.
- Tüm analizlerin sonucunda, Türkçe öğretmeni adaylarının hazırladıkları sınavları değerlendirmede kullanılabilecek bir rubrik oluşturulmuştur.
- Verilerin çözümlenmesi SPSS 24.0 paket programı kullanılarak gerçekleştirilmiştir.

Bulgular

Rubriğin Geçerlik Boyutuna İlişkin Elde Edilen Bulgular

Kapsam geçerliği, bir ölçme aracındaki maddelerin hedeflenen niteliği içerik yönünden ne derece temsil ettiğini ortaya koyan bir göstergedir (Büyüköztürk, 2014). Buna karşılık yapı geçerliği, geliştirilen aracın ölçmeyi amaçladığı kavramı ya da özelliği ne ölçüde doğru yansıttığını inceleyen bir geçerlik türüdür (Seçer, 2015). Türkçe öğretmeni adaylarının sınav hazırlama becerilerini değerlendirmek amacıyla oluşturulan rubrikte, kapsam geçerliğini belirlemek için Lawshe yaklaşımı kullanılmıştır. Yapı geçerliğinin incelenmesinde ise rubrik puanlarının normal dağılım varsayımını karşılamaması nedeniyle puanlayıcılar arası ilişki Spearman sıra farkları korelasyon katsayısı ile analiz edilmiştir. Ayrıca kapsam geçerliği sürecinin sağlıklı yürütülebilmesi için alanda uzman kişilerden görüş alınmış, uzmanlardan her madde için “madde hedeflenen yapıyı ölçüyor”, “madde kısmen ilişkili ancak gereksiz” ya da “madde ilgili yapıyı ölçmüyor” biçiminde değerlendirme yapmaları istenmiştir. Bu uzman değerlendirmeleri, maddelerin içerik açısından uygunluğunu belirlemede önemli bir temel oluşturmıştır.

Her bir maddenin kapsam geçerlik oranını (KGO) hesaplamak için öncelikle o maddeyi “gerekli” olarak değerlendiren uzmanların sayısı dikkate alınır. Bu sayı, toplam uzman sayısının yarısıyla karşılaştırılır ve Lawshe’nin önerdiği formül kullanılarak KGO değeri elde edilir. Buna göre uzmanların tam olarak yarısı bir maddeyi “gerekli” olarak işaretlemişse KGO değeri 0 çıkar. “Gerekli” diyenlerin sayısı yarıdan fazla olduğunda KGO pozitif; yarıdan az olduğunda ise KGO negatif olur. Negatif ya da sıfır değere sahip maddeler, kapsam geçerliği açısından yetersiz kabul edilir. Tüm maddeler için hesaplanan bu oranların ortalaması alınarak kapsam geçerlik indeksi (KGİ) elde edilir. KGİ, ölçme aracındaki maddelerin hedeflenen beceriyi ne ölçüde temsil ettiğini genel düzeyde gösteren bir değerdir. Lawshe’nin (1975) yaklaşımına göre kapsam geçerliği, bir maddenin alana katkısı konusunda uzman görüşlerine dayalı olarak değişiklik gösteren ve maddenin gerekliliğine ilişkin ortak uzlaş

düzeyini yansıtan bir ölçüttür. KGO'nun hesaplanmasında temel ilke, bir maddeyi “gerekli” olarak nitelendiren uzman sayısının, toplam uzman sayısının yarısına oranlanması ve bu değerden 1 çıkarılmasıdır.

Tablo 2. *Kapsam Geçerlik Oranı (KGO) ve İndeksi (KGİ)*

Maddeler	Gerekli	Yararlı/Yetersiz	Gereksiz	Kapsam Geçerlik Oranları (KGO)	Geçerlik
Madde 1	7	0	0	1	
Madde 2	7	0	0	1	
Madde 3	7	0	0	1	
Madde 4	6	1	0	0,71	
Madde 5	7	0	0	1	
Madde 6	7	0	0	1	
Madde 7	6	1	0	0,71	
Uzman Sayısı	7				
Kapsam Geçerlik Ölçütü	0,62				
Kapsam Geçerlik İndeksi (KGİ)	0,917				

Tablo incelendiğinde rubriğin kapsam geçerliğini belirlemek üzere yapılan Lawshe analizinde her bir maddenin uzmanlar tarafından nasıl değerlendirildiği görülmektedir. Yedi uzmandan oluşan grup, maddelerin büyük çoğunluğunu “gerekli” olarak işaretlemiştir. Madde 1, 2, 3, 5 ve 6 için bütün uzmanların “gerekli” seçeneğini tercih etmesi sonucunda bu maddelerin kapsam geçerlik oranı 1,00 olarak hesaplanmıştır. Madde 4 ve Madde 7 için ise altı uzman “gerekli”, bir uzman “yararlı/yetersiz” değerlendirmesinde bulunmuş; bu maddelerin KGO değeri 0,71 olarak belirlenmiştir. Lawshe tablosuna göre yedi uzmanla yapılan değerlendirmelerde bir maddenin kapsam geçerli sayılabilmesi için en az 0,62 değerine ulaşması gerekmektedir. Bu eşik dikkate alındığında rubriğin tüm maddelerinin kapsam geçerlik ölçütünü karşıladığı anlaşılmaktadır. Ayrıca maddelere ait KGO değerleri üzerinden hesaplanan kapsam geçerlik indeksi (KGİ) 0,917 olarak bulunmuş, bu sonuç da rubriğin kapsam geçerliğinin yüksek olduğunu göstermektedir.

Rubriğe ilişkin puanlayıcılar arası ilişkiyi incelemek amacıyla Spearman korelasyon katsayısı hesaplanmıştır. Korelasyon katsayısının 1,00'e yaklaşması puanlayıcılar arasında yüksek düzeyde ilişkiyi, 0,00'a yaklaşması ise ilişkinin zayıf olduğunu göstermektedir. Bu doğrultuda puanlayıcılar tarafından verilen puanlar arasındaki ilişki düzeyi her bir ölçüt ve toplam puan üzerinden ayrı ayrı incelenmiştir. Güvenirlilik incelemesinde ise iki puanlayıcının verdiği puanların tutarlılığını belirlemek amacıyla Cronbach's Alpha (α) iç tutarlılık katsayısı hesaplanmıştır. Bunun yanında puanlayıcılar arasındaki uyum düzeyini ortaya koymak amacıyla Cohen'in Kappa katsayısı da kullanılmıştır. Elde edilen bulgular, rubriğin puanlayıcılar arasında tutarlı sonuçlar üretebildiğini ve güvenilirliğine ilişkin olumlu kanıtlar sunduğunu göstermektedir. Tüm istatistiksel işlemler SPSS 24.0 yazılımı kullanılarak gerçekleştirilmiştir.

Tablo 3. *Puanlayıcılar Arası Spearman Korelasyon Katsayısı*

Maddeler	ρ (rho)	p	N	Yorum
Madde 1	,897	< ,001	50	Çok yüksek ilişki
Madde 2	,817	< ,001	50	Yüksek ilişki
Madde 3	,900	< ,001	50	Çok yüksek ilişki
Madde 4	,708	< ,001	50	Yüksek ilişki
Madde 5	,814	< ,001	50	Yüksek ilişki
Madde 6	,885	< ,001	50	Çok yüksek ilişki
Madde 7	,900	< ,001	50	Çok yüksek ilişki
Toplam	,965	< ,001	50	Çok yüksek ilişki

Bu çalışmada puanlayıcıların madde ve toplam puanlarına ilişkin dağılım özelliklerini değerlendirmek amacıyla normallik testleri uygulanmıştır. Örneklem büyüklüğünün 50 olması nedeniyle normalliğin belirlenmesinde Shapiro-Wilk testi tercih edilmiştir. Bu test, küçük ve orta düzeydeki örneklerde Kolmogorov-Smirnov testine göre daha duyarlı kabul edildiği için literatürde önerilen yöntemdir. Yapılan analizlerde, iki puanlayıcının madde puanlarının normal dağılım göstermediği belirlenmiş ($p < ,05$) ve bu sonuca dayanarak puanlayıcılar arası ilişkiyi incelemeye parametrik olmayan Spearman sıra farkları korelasyon katsayısının kullanılması uygun görülmüştür. Spearman korelasyon katsayıları hem madde düzeyinde hem de toplam puan düzeyinde anlamlı bulunmuş ($p < ,001$) ve ,708 ile ,965 arasında değişmiştir. Bu sonuçlar, rubriğin puanlayıcılar arasında yüksek düzeyde ilişki sağladığını göstermektedir.

Rubriğin Güvenirlik Boyutuna İlişkin Elde Edilen Bulgular

Türkçe öğretmeni adaylarının hazırladığı sınavların değerlendirilmesinde kullanılan rubriğin güvenilirliğini incelemek amacıyla iki bağımsız puanlayıcıdan elde edilen veriler üzerinde Cronbach's Alpha iç tutarlılık katsayısı ile Cohen'in Kappa uyum katsayısı hesaplanmıştır. Cronbach's Alpha katsayısı, ölçme aracında yer alan maddelerin aynı yapıyı ölçüp ölçmediğini değerlendiren bir göstergedir ve maddelerin varyansları ile toplam varyans arasındaki ilişkiye dayalı olarak hesaplanır. Bu katsayının 0,00 ile 1,00 arasında değiştiği, 0,00-0,40 aralığındaki değerlerin güvenirliliğin düşük olduğunu, 0,40-0,60 aralığındaki değerlerin sınırlı güvenirliliğe işaret ettiğini, 0,60-0,80 arasındaki değerlerin ölçme aracının oldukça güvenilir olduğunu ve 0,80-1,00 aralığındaki değerlerin ise yüksek düzeyde güvenirliliği yansıttığı belirtilmektedir (Kalaycı, 2010). Rubriğe ilişkin Cronbach's Alpha sonuçları Tablo 4'te detaylandırılmıştır.

Tablo 4. *Rubriğin İç Tutarlılığına İlişkin Cronbach's Alpha Sonuçları*

N	Cronbach Alpha Katsayısı
50	,772

Rubriğin iç tutarlılığını belirlemek için gerçekleştirilen Cronbach's Alpha analizi, ölçeğin tüm maddelerini kapsayan yapının Alpha katsayısının ,772 olduğunu ortaya koymuştur. Bu değer, alan yazında kabul edilen ,70 sınırının üzerinde olduğu için rubriğin genel anlamda güvenilir bir iç tutarlılık sergilediğini göstermektedir. Ölçeğin bir bütün olarak ele alınması, maddelerin aynı yapıyı yeterli düzeyde yansıttığını ve rubriğin değerlendirme sürecinde tutarlı bir ölçme işlevi sunduğunu ortaya koymaktadır. Elde edilen bulgular, rubrik maddelerinin bütüncül kullanımında güvenilir bir yapı oluşturduğunu doğrulamaktadır.

Cohen'in Kappa katsayısı, rubriklerde iki bağımsız puanlayıcının verdiği puanlar arasındaki uyumu değerlendirmede kullanılan güvenilir istatistiksel yöntemlerden biridir (Cohen, 1960). Bu katsayı, puanlayıcıların alt kategorilerde ne ölçüde aynı değerlendirmeleri yaptığını ortaya koyarken ayrıca görünen uyumun ne kadarının tesadüfi olmaktan kaynaklandığını da hesaba katar. Bu yönüyle, yalnızca yüzdelik hesaplara dayanan uyum ölçümlerinden daha sağlam sonuçlar sunduğu kabul edilmektedir. Kappa katsayısı 0,00 ile 1,00 arasında değişmekte olup değer 1,00'e yaklaşması puanlayıcılar arasındaki tutarlılığın güçlendiğini göstermektedir (Kutlu, Doğan ve Karakaya, 2014). Cohen's Kappa katsayısının yorumlanmasında aşağıdaki sınıflandırma esas alınmıştır: 0'ın altındaki değerler şansa bağlı uyumdan daha düşük bir uyuma işaret ederken, 0,01-0,20 arası önemsiz düzeyde uyumu, 0,21-0,40 arası zayıf düzeyde uyumu, 0,41-0,60 arası orta düzeyde uyumu, 0,61-0,80 arası iyi düzeyde uyumu ve 0,81-1,00 arası ise çok iyi düzeyde uyumu ifade etmektedir (Kılıç, 2015). Bu çalışmada elde edilen Cohen's Kappa değerleri Tablo 5'te ayrıntılı olarak sunulmuştur:

Tablo 5. Rubrik Maddelerine İlişkin Puanlayıcılar Arası Cohen's Kappa Katsayıları

Maddeler	Kappa	p	N	Uyum Düzeyi
Madde 1	,657	< ,001	50	İyi
Madde 2	,641	< ,001	50	İyi
Madde 3	,735	< ,001	50	İyi
Madde 4	,730	< ,001	50	İyi
Madde 5	,639	< ,001	50	İyi
Madde 6	,730	< ,001	50	İyi
Madde 7	,735	< ,001	50	İyi

Cohen's Kappa katsayılarına ilişkin bulgular, iki puanlayıcı arasında iyi düzeyde bir uyum bulunduğunu göstermektedir. Maddelere ait Kappa değerlerinin ,639 ile ,735 arasında değişmesi, puanlayıcıların değerlendirmelerinde benzer ölçütleri esas aldığını ortaya koymaktadır. Tüm değerlerin ,60'ın üzerinde olması, rubrik maddelerinin farklı puanlayıcılar tarafından tutarlı biçimde yorumlandığını ve puanlama sürecinde önemli bir farklılaşma olmadığını göstermektedir. Bu durum, rubriğin değerlendirme sürecinde puanlayıcılar arası uyumu destekleyen bir yapı sunduğuna işaret etmektedir.

Sonuç

Bu araştırma, Türkçe öğretmeni adaylarının sınav hazırlama süreçlerindeki yeterliklerini nesnel biçimde inceleyebilmek amacıyla geliştirilen analitik rubriğin niteliklerini ortaya koymayı amaçlamıştır. Öğretmen adaylarının ölçme ve değerlendirme alanındaki bilgi ve becerileri hem sınıf içi uygulamaların niteliğini hem de öğrenci başarısının izlenmesini doğrudan etkileyen kritik unsurlardan biridir. Bu bağlamda sınav hazırlama sürecindeki kararların sistematik biçimde değerlendirilmesini sağlayan kullanılabilir bir ölçme aracına duyulan gereksinim çalışmanın temel çıkış noktası olmuştur. Rubriğin geliştirilmesi sırasında alan yazında önerilen bilimsel ilkelere uyulmuş, ölçütler açık biçimde tanımlanmış ve içerik yapılandırılmış bir bütün hâline getirilmiştir.

Uzman görüşlerine dayalı yürütülen kapsam geçerliği analizleri, rubrik maddelerinin alanı temsil etme düzeyinin yüksek olduğunu göstermektedir. Uzmanların büyük çoğunluğunun maddeleri "gerekli" olarak değerlendirmesi, ölçütlerin Türkçe öğretimi bağlamındaki sınav hazırlama sürecine uygun biçimde seçildiğini ortaya koymaktadır. Bu bulgu, rubriğin içerik yönünden tutarlı ve öğretmen

adaylarının değerlendirmesi amacıyla kullanılabilir bir çerçeve sunduğunu göstermesi bakımından önemlidir.

Puanlayıcılar arası ilişkiye dayalı bulgular, rubriğin farklı değerlendiriciler tarafından benzer biçimde uygulanabildiğini göstermiştir. İki puanlayıcıdan elde edilen puanlar arasındaki yüksek düzeyde ilişki, değerlendirme ölçütlerinin açık ve işlevsel olduğunu, soruların puanlanmasında benzer kriterlerin kullanıldığını ortaya koymaktadır. Korelasyon katsayılarının anlamlı ve güçlü olması, öğretmen adaylarının sınav tasarlama becerilerinin rubrik aracılığıyla tutarlı biçimde değerlendirilebildiğini göstermektedir. Bu sonuç, rubriğin ölçme amacıyla kullanılabilir bir yapı sunduğunu desteklemektedir.

Güvenirlilik analizlerinden elde edilen bulgular da aracın bütüncül niteliğini ortaya koymaktadır. İç tutarlılık katsayısının kabul edilebilir düzeyin üzerinde olması, maddelerin aynı yapıyı ölçtüğüne ve değerlendirme sürecinin içsel olarak tutarlı işlediğine işaret etmektedir. Ayrıca puanlayıcılar arasındaki uyumu gösteren katsayıların yüksek bulunması, rubriğin farklı değerlendiriciler tarafından benzer sonuçlar üretebildiğini ortaya koymakta ve aracı puanlayıcı bağımsızlığı açısından güçlü kılmaktadır.

Bu çalışmada geliştirilen rubrik, öğretmen adaylarının sınav hazırlama sürecine ilişkin performanslarını tanımlayıcı bir bakış açısıyla değerlendirerek hem öğretim elemanlarına hem de adaylara önemli bir geri bildirim olanağı sunmaktadır. Öğretim elemanları bu rubrik aracılığıyla adayların sınav hazırlarken hangi aşamalarda zorlandıklarını, hangi ölçütleri başarıyla yerine getirdiklerini veya hangi alanlarda desteklenmeleri gerektiğini belirleyebilecektir. Öğretmen adayları açısından ise rubrik, sınav hazırlama sürecini daha bilinçli, planlı ve hedefe uygun biçimde yürütmeye katkı sağlayacaktır. Özellikle beceri temelli ölçme anlayışının öne çıktığı güncel programlar dikkate alındığında bu rubriğin hem öğretmen eğitimi sürecinde hem de uygulamalı araştırmalarda kullanılabilirliği yüksektir.

Genel olarak değerlendirildiğinde çalışma sonucunda geliştirilen analitik rubrik; kapsam geçerliğine ilişkin bulgular, puanlayıcılar arası ilişki düzeyi ve güvenirlilik analizleri birlikte ele alındığında, Türkçe öğretmeni adaylarının sınav hazırlama becerilerini değerlendirmede kullanılabilir ve tutarlı sonuçlar üretebilen bir ölçme aracı niteliği taşımaktadır. Bu aracın öğretmen yetiştirme programlarına ve ileride yapılacak araştırmalara katkı sağlayacağı düşünülmektedir.

Çatışma Beyanı ve Etik Bildirim

Araştırma ile ilgili diğer kişi ve kurumlarla yaşanabilecek herhangi bir çıkar çatışması yoktur. Etik kurul onayı, Zonguldak Bülent Ecevit Üniversitesi İnsan Araştırmaları Etik Kurulundan 15.12.2025 tarihli ve 685172 sayılı kararı ile alınmıştır.

Kaynaklar

- Andrade, H. G. (1997). Understanding rubrics. *Educational Leadership*, 54(4), 14–17.
- Büyüköztürk, Ş. (2014). *Sosyal bilimler için veri analizi el kitabı*. Pegem Akademi.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46.
- Çiçek, S. (2024). Ana dili olarak Türkçe öğretiminde dört temel dil becerisini ölçme ve değerlendirmeye ilişkin öğretmen görüşleri: Beceri temelli yazılı ve uygulamalı sınav sistemi. *Dokuz Eylül Üniversitesi Buca Eğitim Fakültesi Dergisi*, 62, 3517–3542.

- Doğan Kahtalı, B., & Çelik, Ş. (2019). 2019 Türkçe öğretim programında ölçme ve değerlendirme ile Türkçe öğretmenlerinin ölçme ve değerlendirme araçlarını kullanma düzeyleri. *Eğitim Kuram ve Uygulama Araştırmaları Dergisi*, 6(2), 237–244.
- Kalaycı, Ş. (2010). *SPSS uygulamalı çok değişkenli istatistik teknikleri*. Asil Yayın Dağıtım.
- Karaman, P. (2022). Ölçme-değerlendirme okuryazarlığı. M. Özden (Ed.), *Okur yazar mıyım* (pp. 93–105) içinde. Anı Yayıncılık.
- Karatay, H., & Dilekçi, A. (2019). Türkçe öğretmenlerinin dil becerilerini ölçme ve değerlendirme yeterlikleri. *Millî Eğitim Dergisi*, 48(1), 685–716.
- Khan, R. (2023). Exploring assessment literacy of tertiary-level teachers in Bangladesh. In *Local research and global perspectives in English language teaching: Teaching in changing times* (pp. 345–361). Springer Nature Singapore.
- Kintsch, E. (2005). Comprehension theory as a guide for the design of thoughtful questions. *Topics in Language Disorders*, 25, 51–64.
- Kılıç, S. (2015). Kappa testi. *Journal of mood disorders*, 5(3), 142–144.
- Kolcuoğlu, B., & Boylu, E. (2024). Yabancı dil olarak Türkçe öğrenenlerin ölçme ve değerlendirme okuryazarlıkları üzerine bir inceleme. *Türkiye Sosyal Araştırmalar Dergisi*, 28(1), 249–272.
- Kutlu, Ö., Doğan, C. D., & Karakaya, İ. (2014). *Ölçme ve değerlendirme: Performansa ve portfolyoya dayalı durum belirleme*. Pegem Akademi.
- Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology*, 28, 563–575.
- Mertler, A. C. (2004). Secondary teachers' assessment literacy: Does classroom experience make a difference? *American Secondary Education*, 33(1), 49–64.
- MEB (2024). Türkiye yüzyılı maarif modeli ortaokul Türkçe dersi öğretim programı.
- Sallabaş, M. E., & Kiler, B. (2024). Dil becerilerini ölçme ve değerlendirme sürecinde Türkçe öğretmenlerinin karşılaştıkları sorunlar ve çözüm önerileri. *Ana Dili Eğitimi Dergisi*, 12(3), 650–666.
- Seçer, İ. (2015). *Psikolojik test geliştirme ve uyarlama süreci: SPSS ve LISREL uygulamaları*. Anı Yayıncılık.
- Stiggins, R. J. (1995). Assessment literacy for the 21st century. *Phi Delta Kappan*, 77(3), 238–245.
- Türkben, T. (2022). Türkçe öğretmenlerinin sınıf içi öğrenme becerilerini ölçme ve değerlendirme yeterlikleri. *Eğitim ve İnsani Bilimler Dergisi: Teori ve Uygulama Dergisi*, 13(25), 45–72.
- Ömeroğlu, A. F. (2024). Türkçe öğretmenlerinin uyguladığı sınav sorularının öğrenme alanları, ölçme değerlendirme yöntemleri ve bilişsel alan basamakları açısından incelenmesi. *Korkut Ata Türkiyat Araştırmaları Dergisi*, 14, 975–992.
- Ünlü, S., Öztürk, H., & Tağa, T. (2014). Türkçe dersinde uygulanan sınavlar üzerine bir değerlendirme. *International Journal of Social Science*, 28, 513–523.
- YÖK (2018). Öğretmen yetiştirme lisans programları. <http://www.yok.gov.tr>
- Yüceer, D. (2024). Türkçe öğretmeni adaylarının soru hazırlamaya dair deneyim ve görüşleri. *Ana Dili Eğitimi Dergisi*, 12(1), 69–86.

RESEARCH

Open Acces

**A Study on Developing a Rubric to Measure Exam Preparation
Skills of Prospective Turkish Language Teachers**

Efecan Karagöl *

Abstract

This research aims to develop a comprehensive analytical rubric to systematically, objectively, and measurably assess the exam preparation and evaluation skills of prospective Turkish language teachers. Recognizing that pre-service teachers' measurement and evaluation knowledge directly impacts both instructional quality and student achievement, a holistic and structured assessment tool is critically essential. The rigorous development process of this rubric encompassed a detailed literature review, criteria establishment, draft form creation, and expert-led content validation. According to Lawshe's analysis, all generated items successfully satisfied the content validity criteria. Furthermore, the inter-rater reliability of the rubric was thoroughly examined using the Spearman correlation coefficient, with values found to vary between .708 and .965. In the context of reliability analyses, the Cronbach's Alpha coefficient was calculated as .772, indicating that the rubric demonstrates a highly satisfactory degree of internal consistency. Additionally, Cohen's Kappa coefficients demonstrated a substantial level of inter-rater agreement across all items. Ultimately, the findings obtained prove that the developed analytical rubric is a valid, reliable, and highly viable measurement tool for evaluating the exam preparation and assessment skills of pre-service Turkish teachers.

Keywords: Exam preparation, rubric development, pre-service Turkish teachers, analytical rubric, validity, reliability.

Article History

Received: 13/02/2026

Correction: 19/03/2026

Accept: 09/05/2026

* Prof. Dr., Zonguldak Bülent Ecevit University, 0000-0003-0331-8009, Zonguldak, Türkiye, efecankaragol@gmail.com

For citation: Karagöl, E. (2026). A study on developing a rubric to measure exam preparation skills of prospective Turkish language teachers. *OJCES*, 4(1), 87-108. <https://doi.org/10.5281/zenodo.20802678>

Introduction

The competencies that pre-service Turkish teachers must acquire during the exam preparation process are of paramount importance in fostering competent educational environments. It is imperative that these competencies are reinforced through quantifiable and modifiable methodologies that extend beyond mere theoretical understanding. Within this framework, there is a need for an organized and reliable assessment tool that can reveal the measurement and evaluation competencies of pre-service teachers. Conversely, rubrics are regarded as efficacious instruments that can address this necessity by offering the capacity to assess individuals' performances in accordance with predefined criteria.

The efficacy with which an educator conducts a teaching process is directly associated with their capacity to formulate high-quality inquiries that prompt students to engage in critical thinking, comprehend the material, and apply their knowledge (Yüceer, 2024). In this context, a teacher's capacity to execute evaluation processes congruent with their intended purpose necessitates meticulous planning, systematic structuring, and systematic implementation of said processes to ensure alignment with the objectives of the designated learning area. The questions employed in the classroom serve a dual purpose: they reveal the students' current knowledge levels and contribute to their construction of meaning, thereby deepening the learning process. In this functional aspect, questions serve two fundamental functions: determining the extent to which students remember, comprehend, and can apply what has been taught (Kintsch, 2005). Furthermore, effective questions have been demonstrated to activate students' higher-order thinking skills, thereby laying the groundwork for critical and creative thinking. These tools empower educators by enabling them to orchestrate classroom interactions, discern students' needs, and adapt the teaching process in real-time. At this juncture, teachers' measurement and evaluation literacy becomes a critical necessity for two reasons. First, it can help them develop more effective and conscious measurement and evaluation processes. Second, it can help them create better questions.

Although measurement and evaluation literacy is considered one of the core components of teachers' professional competencies, it has been observed that this field is often overlooked in teacher education programs (Khan, 2023, p. 345). However, this literacy, which encompasses the knowledge and skills to discuss, implement, and design measurement and evaluation strategies, is an indispensable element of an effective teaching process (Mertler, 2004). Indeed, Stiggins (1995) emphasizes that teachers who are literate in measurement and evaluation should have sufficient equipment in critical areas such as the evaluation of students' knowledge and abilities, the determination of the most effective methods for this purpose, the identification of factors that negatively affect measurement processes, and the prevention of these factors. The absence of these competencies hinders teachers' ability to demonstrate the expected level of performance in their evaluations, thereby diminishing the efficacy of their pedagogical practices (Kolcuoğlu & Boylu, 2024). A teacher who possesses the necessary competencies in measurement and evaluation is able to formulate suitable measurement and evaluation activities to ascertain students' learning outcomes. These activities are then carried out in an effective manner, and feedback of a high caliber is provided to support student development based on the results obtained. In this context, the ability of the teacher to integrate their knowledge and skills in the field of measurement and evaluation with their pedagogical knowledge, content knowledge, and the classroom learning environment is a fundamental condition for successfully fulfilling both the "instructor" and "evaluator" roles (Karaman, 2022). These integrated competencies unequivocally illustrate the necessity of meticulous attention to the elements to be considered in the process of exam preparation.

In the context of Turkish teacher training, while measurement and evaluation knowledge and skills are accorded a significant place in the overall program structure, an examination of course contents reveals that this field is often addressed within a fragmented and theoretical framework. The "Measurement and Evaluation in Education" course, a fundamental component of the curriculum, encompasses foundational principles pertaining to measurement and evaluation, the psychometric properties of measurement instruments, the development process of achievement tests, item analysis, interpretation of results, and technical dimensions such as grading (CoHE/YÖK, 2018). Conversely, the elective course "Exam Preparation and Evaluation in Turkish Language Teaching" addresses exam types for different age groups and language levels, principles for measuring language skills, question types regarding reading-writing-listening-speaking and vocabulary/grammar, exam preparation techniques, and evaluation criteria (CoHE/YÖK, 2018). Conversely, an evaluation of the contents of both courses reveals that pre-service teachers are not provided with a systematic structure that would integrate all this information into a holistic exam design process, thereby enabling them to experience content validity and item quality in a practical manner. This deficiency complicates the integration of psychometric principles and the knowledge of question-writing expertise specific to language skills into a coherent exam preparation process, resulting in uncertainties and practical inadequacies in exam design. Consequently, this scenario necessitates the formulation of a competent rubric that can objectively assess the exam preparation performance of pre-service teachers and elucidate their areas for enhancement.

The 2024 Secondary School Turkish Language Curriculum places significant emphasis on the integration of measurement and evaluation as inseparable components of the teaching process. The curriculum delineates that the fundamental objective of classroom assessment is to facilitate the learning process. Consequently, it is imperative that measurement tools be designed with the following characteristics: clarity, comprehensibility, appropriateness for the student's level, and the inclusion of diverse evaluation tasks. Additionally, the curriculum underscores the importance of task formats in summative assessments that are novel to students, as these formats can potentially lead to failure. Consequently, evaluation processes should be founded on the principles of continuity and familiarity. The program delineates measurement types for reading, listening, speaking, and writing skills. It stipulates that evaluation should be conducted using checklists and rubrics after production tasks, and that high-quality, process-oriented feedback should be provided to the student (MoNE/MEB, 2024). This framework underscores the teacher's dual responsibility: not only to formulate items but also to establish an evaluation system that is consistent with learning outcomes, skill-based, process-oriented, and based on clear criteria. However, given the absence of practical standards and sample exam templates that illustrate the integration of these principles into a concrete exam preparation process within the curriculum, significant challenges arise, particularly in the context of teacher education, to systematically develop and monitor exam preparation competencies. This situation underscores the necessity for tools to assist pre-service teachers in transforming measurement and evaluation principles into a holistic exam design.

A review of the extant literature reveals the absence of a measurement tool in the literature directly aimed at measuring the exam preparation and evaluation skills of pre-service Turkish teachers. This is despite the existence of studies addressing pre-service teachers' views on exam preparation and measurement-evaluation (Çiçek, 2024; Sallabaş & Kiler, 2024; Yüceer, 2024). Furthermore, research examining measurement-evaluation competencies (Doğan Kahtalı & Çelik, 2019; Karatay & Dilekçi, 2019; Türkben, 2022) and studies analyzing exam contents (Ömeroğlu, 2024; Ünlü, Öztürk & Taha, 2014)

also exist. This discrepancy complicates the objective, systematic, and comparable assessment of pre-service teachers' competencies in exam design processes. The objective of this research is to develop a valid and reliable analytical rubric that can assess the exam preparation and evaluation skills of pre-service Turkish teachers.

Method

Research Model

The development of the rubric was informed by the contributions of three distinct participant groups. To ensure content validity, an expert group consisting of academicians with the titles of Professor, Associate Professor, and Assistant Professor working in various provinces of Türkiye was included in the process. The characteristics of the participants involved in the study are presented in Table 1.

Table 1. *Characteristics of the Expert Group*

	Asst. Prof.	Assoc. Prof.	Prof.
Female	2	1	-
			1
Gender Male	1	2	
Total	3	3	1

As demonstrated in Table 1, the study encompassed Assistant Professors (f=3), Associate Professors (f=3), and a Professor (f=1). To assess the construct validity and reliability of the research, a sample of 50 fourth-year pre-service teachers enrolled in the Department of Turkish Language Education at a university in the Western Black Sea region were selected for participation in the study. The subjects were selected using the convenience sampling method. Furthermore, within the scope of the reliability dimension of the study, simultaneous scoring was performed by the researcher and one expert to determine inter-rater agreement.

Data Collection and Analysis

In the development process of the rubric designed to measure the exam preparation and evaluation skills of pre-service Turkish teachers, the fundamental stages proposed by Andrade (1997) were taken into consideration. Accordingly, the relevant literature was first comprehensively reviewed, followed by the determination of the evaluation criteria and performance levels. In accordance with the aforementioned elements, a preliminary rubric was formulated and subsequently implemented. In light of the feedback obtained during the implementation process, necessary revisions were made to the rubric. After these revisions, validity and reliability analyses were conducted. Following the culmination of these stages, the rubric was finalized into its ultimate form as a viable assessment instrument. The research process, aligned with Andrade's recommendations, is illustrated in Figure 1:

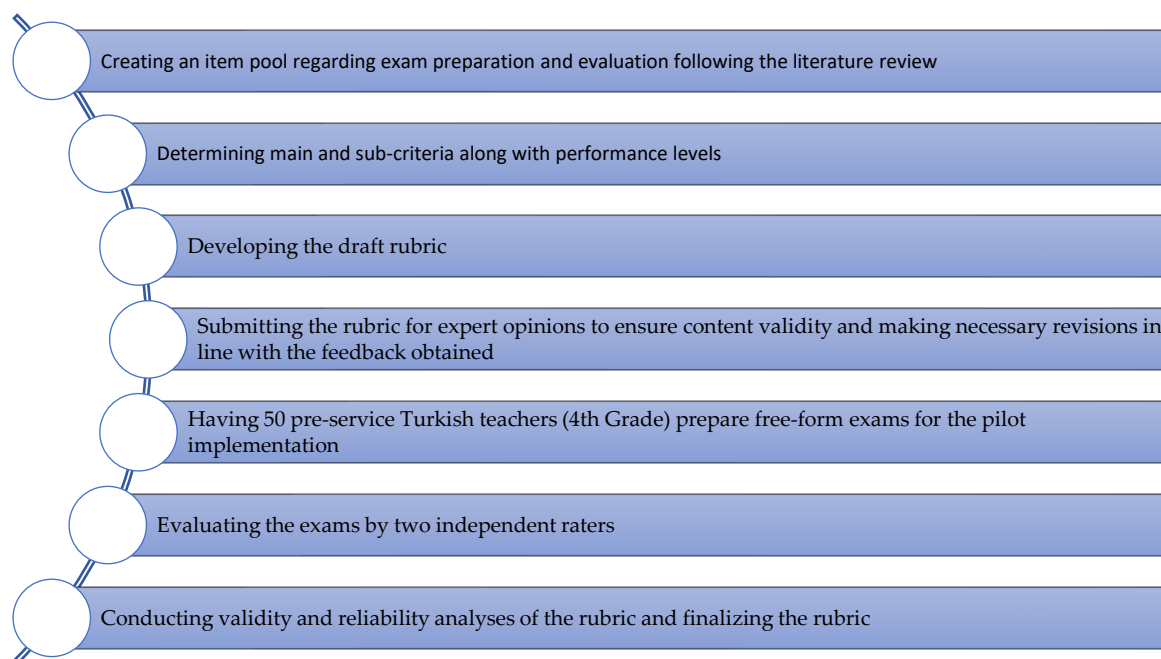


Figure 1. Rubric development process.

- As illustrated in Figure 1, a comprehensive item pool was meticulously constructed to assess the exam preparation skills of pre-service Turkish teachers following a thorough literature review.
- The rubric developed within the scope of this study is predicated on two fundamental dimensions. The competencies addressed in this text include "exam design skills" and "evaluation and objectivity skills." The "exam design skills" dimension comprises the following sub-criteria: alignment of skill areas with learning outcomes, question variety, balanced distribution of questions across cognitive levels, compliance with language and expression rules, and instructions. The "evaluation and objectivity skills" dimension encompasses the sub-criteria of evaluation criteria and objectivity. Each criterion in the rubric is scored across four levels: The following scale was used to evaluate the quality of the submissions: "Good (4)," "Fair (3)," "Needs Improvement (2)," and "Inadequate (1)." According to this structure, the total score range that can be obtained from an exam product varies between 7 and 28. In the interpretation of the scores obtained from the rubric, both the total scores and the average scores of the criteria were considered. As each criterion is scored on a scale ranging from 1 to 4, interpretations were derived based on the mean scores obtained during the evaluation process. The interval coefficient was utilized as the foundation for determining the average score ranges, and performance levels were classified accordingly.
- Table 2: The performance levels are presented in accordance with the average scores. Accordingly, the range of 1.00–1.75 is designated as "inadequate," 1.76–2.50 as "needs improvement," 2.51–3.25 as "fair," and 3.26–4.00 as "good." The interval coefficient was considered in determining these average score ranges. The interval coefficient was determined through the calculation of the ratio between the maximum and minimum scores, divided by the total number of performance levels. This calculation yielded a value of 0.75.
- The performance levels were determined based on the average score ranges, as outlined below:
 - 1.00–1.75: Inadequate;
 - 1.76–2.50: Needs improvement;
 - 2.51–3.25: Fair;
 - 3.26–4.00: Good.

- This classification was employed to facilitate the interpretation of the average scores obtained from the rubric.
- The Lawshe method was employed to determine the content validity of the rubric. In this method, a table of specifications is presented to experts, and they are asked to indicate whether or not each item measures the targeted construct. The items are then classified as "the item measures the targeted construct," "the item is partially related but unnecessary," or "the item does not measure the relevant construct."
- In accordance with the expert feedback received, the necessary revisions to the rubric items were made. Subsequent to this, the Content Validity Ratio (CVR) and the Content Validity Index (CVI) were calculated.
- In the course of the pilot implementation process, an exam preparation task was assigned to 50 pre-service Turkish teachers who were in their fourth year and had not yet taken the "Exam Preparation and Evaluation in Turkish Teaching" course.
- The evaluation of these exams was conducted by two independent evaluators (R1 and R2) to ascertain inter-rater consistency.
- The Spearman correlation coefficient was employed to assess the inter-rater reliability concerning the rubric.
- In the course of the reliability analyses, the internal consistency between the scores provided by the two raters was determined using the Cronbach's Alpha coefficient. Furthermore, Cohen's Kappa coefficient was calculated to ascertain the inter-rater agreement.
- As a result of all analyses, a rubric that can be used to evaluate the exams prepared by pre-service Turkish teachers was established.
- Data analysis was performed using the SPSS 24.0 software package.

Findings

Findings Regarding the Validity Dimension of the Rubric

Content validity is a metric that quantifies the extent to which the items in a measurement tool accurately reflect the targeted quality in terms of content (Büyüköztürk, 2014). Construct validity, on the other hand, is a form of validity that examines the extent to which the developed instrument accurately reflects the concept or trait it aims to measure (Seçer, 2015). In the rubric created to evaluate the exam preparation skills of pre-service Turkish teachers, the Lawshe approach was used to determine content validity. In the examination of construct validity, given that the rubric scores did not meet the assumption of normal distribution, the inter-rater relationship was analyzed using the Spearman rank-order correlation coefficient. In addition, to guarantee the effective execution of the content validity process, opinions were obtained from experts in the field. These experts were asked to evaluate each item as "the item measures the targeted construct," "the item is partially related but unnecessary," or "the item does not measure the relevant construct." These expert evaluations constituted a significant foundation for assessing the suitability of the items in terms of content.

The calculation of the content validity ratio (CVR) for each item is primarily determined by the number of experts who evaluate that item as "essential." This number is then compared with half of the total number of experts, and the CVR value is obtained using the formula proposed by Lawshe. Consequently, if precisely half of the experts designate an item as "essential," the CVR value is equivalent to 0. Conversely, when the proportion of experts labeling an item as "essential" exceeds half, the CVR assumes a positive value. Conversely, when the proportion of experts labeling an item as "essential" falls short of half, the CVR assumes a negative value. Items exhibiting negative or zero values are deemed inadequate from the perspective of content validity. The content validity index (CVI) is

obtained by taking the average of these ratios calculated for all items. The CVI is a value that indicates the extent to which the items in the measurement tool represent the targeted skill on a general level. According to Lawshe's (1975) approach, content validity is a criterion that varies based on expert opinions regarding an item's contribution to the field and reflects the level of common consensus regarding the necessity of the item. The fundamental principle in calculating the CVR is to establish a ratio between the number of experts who categorize an item as "essential" and the total number of experts. Then, the value is reduced by one.

Table 2. *Content Validity Ratio (CVR) and Index (CVI)*

Items	Essentials	Useful/Inadequate	Unnecessary	Content Validity Ratios (CVR)
Item 1	7	0	0	1
Item 2	7	0	0	1
Item 3	7	0	0	1
Item 4	6	1	0	0,71
Item 5	7	0	0	1
Item 6	7	0	0	1
Item 7	6	1	0	0,71
Number of Experts	7			
Content Validity Criterion	0,62			
Content Validity Index (CVI)	0,917			

Upon examination of the table, the Lawshe analysis is employed to ascertain the content validity of the rubric, thereby elucidating the evaluation of each item by the experts. The group, which included seven experts, indicated that the majority of the items were deemed "essential." Given the predilection of the experts for the "essential" option with respect to Items 1, 2, 3, 5, and 6, the content validity ratio for these items was determined to be 1.00. For items 4 and 7, six experts provided an "essential" evaluation, while one expert provided a "useful/inadequate" evaluation. The CVR value for these items was determined to be 0.71. According to the Lawshe table, in evaluations conducted with seven experts, an item must reach a value of at least 0.62 to be considered valid in terms of content. Given this established threshold, it is understood that all items of the rubric meet the content validity criterion. Furthermore, the content validity index (CVI) calculated over the CVR values of the items was found to be 0.917, suggesting that the content validity of the rubric is high.

The Spearman correlation coefficient was calculated to examine the inter-rater relationship regarding the rubric. A correlation coefficient approaching 1.00 indicates a high level of relationship between the raters, while a value approaching 0.00 indicates a weak relationship. Consequently, the inter-rater reliability of the scores was examined for each criterion and the total score. In the reliability examination, the Cronbach's Alpha (α) internal consistency coefficient was calculated to determine the consistency of the scores provided by the two raters. Furthermore, Cohen's Kappa coefficient was employed to ascertain the level of agreement among the raters. The findings obtained demonstrate that

the rubric can produce consistent results between raters and present positive evidence regarding its reliability. All statistical procedures were performed using SPSS 24.0 software.

Table 3. *Inter-rater Spearman Correlation Coefficient*

Item / Measure	ρ (rho)	p	N	Interpretation
Item 1	,897	< ,001	50	Very high relationship
Item 2	,817	< ,001	50	High relationship
Item 3	,900	< ,001	50	Very high relationship
Item 4	,708	< ,001	50	High relationship
Item 5	,814	< ,001	50	High relationship
Item 6	,885	< ,001	50	Very high relationship
Item 7	,900	< ,001	50	Very high relationship
Total	,965	< ,001	50	Very high relationship

In this study, normality tests were conducted to evaluate the distribution characteristics of the raters' item and total scores. Given the sample size of 50, the Shapiro–Wilk test was deemed the optimal method for assessing normality. This test is the recommended method in the literature as it is considered more sensitive than the Kolmogorov–Smirnov test for small and medium-sized samples. The subsequent analyses indicated that the item scores of the two raters did not adhere to a normal distribution ($p < .05$). In light of these findings, the non-parametric Spearman's rank correlation coefficient was deemed a suitable metric for assessing the inter-rater relationship. The Spearman correlation coefficients were found to be significant at both the item and total score levels ($p < .001$), ranging between .708 and .965. These results suggest that the rubric exhibits a high degree of inter-rater reliability.

Findings Regarding the Reliability Dimension of the Rubric

The present study sought to examine the reliability of the rubric employed in the evaluation of exams prepared by Turkish teacher candidates. To this end, the Cronbach's Alpha internal consistency coefficient and Cohen's Kappa agreement coefficient were calculated based on data obtained from two independent raters. The Cronbach's Alpha coefficient is a metric that evaluates the degree to which items in a measurement tool measure the same construct. This coefficient is calculated based on the relationship between the variances of the items and the total variance. The coefficient's range is reported to be from 0.00 to 1.00. Values between 0.00 and 0.40 are indicative of low reliability, values between 0.40 and 0.60 indicate limited reliability, values between 0.60 and 0.80 indicate moderate reliability, and values between 0.80 and 1.00 indicate high reliability (Kalaycı, 2010). The Cronbach's Alpha results regarding the rubric are detailed in

Table 4. *Cronbach's Alpha Results Regarding the Internal Consistency of the Rubric*

N	Cronbach's Alpha Coefficient
50	,772

The Cronbach's Alpha analysis was conducted to ascertain the internal consistency of the rubric, and the Alpha coefficient for the structure encompassing all items of the scale was .772. Given that this value exceeds the .70 threshold frequently cited in the extant literature, it can be deduced that the rubric in its entirety evinces reliable internal consistency. On the basis of the entire scale, it is evident that the items reflect the same construct at an adequate level. Furthermore, the rubric functions as a consistent measurement tool in the evaluation process. The findings indicate that the rubric items form a reliable

structure when used holistically. Cohen's Kappa coefficient is a reliable statistical method employed to evaluate the agreement between scores provided by two independent raters in rubrics (Cohen, 1960). While this coefficient reveals the extent to which raters make the same evaluations in subcategories, it also accounts for how much of the observed agreement may be due to chance. In this regard, it is generally accepted to provide more robust results than agreement measurements based solely on percentage calculations. The Kappa coefficient, a measure of agreement between raters, ranges from 0.00 to 1.00. As the value approaches 1.00, it indicates a strengthening consistency between raters (Kutlu, Doğan, & Karakaya, 2014). The following classification was used to interpret the Cohen's Kappa coefficient: values below 0 indicate agreement lower than expected by chance, while 0.01–0.20 indicates slight agreement, 0.21–0.40 indicates fair agreement, 0.41–0.60 indicates moderate agreement, 0.61–0.80 indicates substantial agreement, and 0.81–1.00 indicates almost perfect agreement (K Kılıç, 2015). The Cohen's Kappa values obtained in this study are presented in detail in Table 5.

Table 5: Inter-rater Cohen's Kappa coefficients regarding the rubric items

Items	Kappa	p	N	Level of Agreement
Item 1	,657	<,001	50	İyi
Item 2	,641	<,001	50	İyi
Item 3	,735	<,001	50	İyi
Item 4	,730	<,001	50	İyi
Item 5	,639	<,001	50	İyi
Item 6	,730	<,001	50	İyi
Item 7	,735	<,001	50	İyi

The findings regarding the Cohen's Kappa coefficients indicate that there is a substantial level of agreement between the two raters. The fact that the Kappa values for the items range between .639 and .735 demonstrates that the raters based their evaluations on similar criteria. All values being above .60 shows that the rubric items were interpreted consistently by different raters and that there was no significant divergence in the scoring process. This situation indicates that the rubric offers a structure that supports inter-rater agreement during the evaluation process.

Conclusion

This research aimed to reveal the characteristics of the analytic rubric developed to objectively examine the competencies of Turkish teacher candidates in their exam preparation processes. The knowledge and skills of teacher candidates in the field of measurement and evaluation are critical elements that directly affect both the quality of classroom practices and the monitoring of student achievement. In this context, the need for a usable measurement tool that ensures the systematic evaluation of decisions during the exam preparation process served as the fundamental starting point of the study. During the development of the rubric, scientific principles recommended in the literature were followed, criteria were clearly defined, and the content was structured into a cohesive whole.

Content validity analyses conducted based on expert opinions indicate that the rubric items have a high level of representation of the field. The fact that the vast majority of experts evaluated the items as "essential" demonstrates that the criteria were appropriately selected for the exam preparation process within the context of Turkish language teaching. This finding is significant as it shows that the rubric provides a content-consistent and usable framework for evaluating teacher candidates.

Findings based on inter-rater relationships showed that the rubric can be applied similarly by different evaluators. The high level of relationship between the scores obtained from two raters reveals that the evaluation criteria are clear and functional, and that similar criteria were used in scoring the questions. The significant and strong correlation coefficients demonstrate that the exam design skills of teacher candidates can be consistently evaluated through the rubric. This result supports the fact that the rubric offers a usable structure for measurement purposes.

Findings obtained from reliability analyses also reveal the holistic quality of the tool. The internal consistency coefficient being above the acceptable level indicates that the items measure the same construct and that the evaluation process functions internally consistently. Furthermore, the high coefficients representing inter-rater agreement demonstrate that the rubric can produce similar results across different evaluators, making the tool strong in terms of rater independence.

The rubric developed in this study provides a significant feedback opportunity for both instructors and candidates by evaluating the performances of teacher candidates regarding the exam preparation process from a descriptive perspective. Through this rubric, instructors will be able to determine at which stages candidates struggle while preparing exams, which criteria they successfully fulfill, or in which areas they need support. For teacher candidates, the rubric will contribute to conducting the exam preparation process more consciously, planned, and in line with objectives. Considering current curricula where skill-based measurement approaches come to the fore, the usability of this rubric is high both in the teacher education process and in applied research.

When evaluated overall, the analytic rubric developed as a result of the study -when findings regarding content validity, the level of inter-rater relationship, and reliability analyses are considered together-possesses the quality of a measurement tool that is usable and capable of producing consistent results in evaluating the exam preparation skills of Turkish teacher candidates. It is believed that this tool will contribute to teacher training programs and future research.

Conflict Declaration and Ethical Statement

There is no conflict of interest with other individuals or institutions related to the research. Ethical committee approval was obtained from the Zonguldak Bülent Ecevit University Human Research Ethics Committee with the decision dated December 15, 2025, and numbered 685172.

References

- Andrade, H. G. (1997). *Understanding rubrics*. *Educational Leadership*, 54(4), 14–17.
- Büyüköztürk, Ş. (2014). *Sosyal bilimler için veri analizi el kitabı*. Pegem Akademi.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46.
- Çiçek, S. (2024). Ana dili olarak Türkçe öğretiminde dört temel dil becerisini ölçme ve değerlendirmeye ilişkin öğretmen görüşleri: Beceri temelli yazılı ve uygulamalı sınav sistemi. *Dokuz Eylül Üniversitesi Buca Eğitim Fakültesi Dergisi*, 62, 3517–3542.
- Doğan Kahtalı, B., & Çelik, Ş. (2019). 2019 Türkçe öğretim programında ölçme ve değerlendirme ile Türkçe öğretmenlerinin ölçme ve değerlendirme araçlarını kullanma düzeyleri. *Eğitim Kuram ve Uygulama Araştırmaları Dergisi*, 6(2), 237–244.
- Kalaycı, Ş. (2010). *SPSS uygulamalı çok değişkenli istatistik teknikleri*. Asil Yayın Dağıtım.

- Karaman, P. (2022). Ölçme-değerlendirme okuryazarlığı. M. Özden (Ed.), *Okur yazar mıyım* (pp. 93–105) içinde. Anı Yayıncılık.
- Karatay, H., & Dilekçi, A. (2019). Türkçe öğretmenlerinin dil becerilerini ölçme ve değerlendirme yeterlikleri. *Millî Eğitim Dergisi*, 48(1), 685–716.
- Khan, R. (2023). Exploring assessment literacy of tertiary-level teachers in Bangladesh. In *Local research and global perspectives in English language teaching: Teaching in changing times* (pp. 345–361). Springer Nature Singapore.
- Kintsch, E. (2005). Comprehension theory as a guide for the design of thoughtful questions. *Topics in Language Disorders*, 25, 51–64.
- Kılıç, S. (2015). Kappa testi. *Journal of mood disorders*, 5(3), 142–144.
- Kolcuoğlu, B., & Boylu, E. (2024). Yabancı dil olarak Türkçe öğretmenlerin ölçme ve değerlendirme okuryazarlıkları üzerine bir inceleme. *Türkiye Sosyal Araştırmalar Dergisi*, 28(1), 249–272.
- Kutlu, Ö., Doğan, C. D., & Karakaya, İ. (2014). *Ölçme ve değerlendirme: Performansa ve portfolyoya dayalı durum belirleme*. Pegem Akademi.
- Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology*, 28, 563–575.
- Mertler, A. C. (2004). Secondary teachers' assessment literacy: Does classroom experience make a difference? *American Secondary Education*, 33(1), 49–64.
- MEB (2024). Türkiye yüzyılı maarif modeli ortaokul Türkçe dersi öğretim programı.
- Sallabaş, M. E., & Kiler, B. (2024). Dil becerilerini ölçme ve değerlendirme sürecinde Türkçe öğretmenlerinin karşılaştıkları sorunlar ve çözüm önerileri. *Ana Dili Eğitimi Dergisi*, 12(3), 650–666.
- Seçer, İ. (2015). *Psikolojik test geliştirme ve uyarlama süreci: SPSS ve LISREL uygulamaları*. Anı Yayıncılık.
- Stiggins, R. J. (1995). Assessment literacy for the 21st century. *Phi Delta Kappan*, 77(3), 238–245.
- Türkben, T. (2022). Türkçe öğretmenlerinin sınıf içi öğrenme becerilerini ölçme ve değerlendirme yeterlikleri. *Eğitim ve İnsani Bilimler Dergisi: Teori ve Uygulama Dergisi*, 13(25), 45–72.
- Ömeroğlu, A. F. (2024). Türkçe öğretmenlerinin uyguladığı sınav sorularının öğrenme alanları, ölçme değerlendirme yöntemleri ve bilişsel alan basamakları açısından incelenmesi. *Korkut Ata Türkiyat Araştırmaları Dergisi*, 14, 975–992.
- Ünlü, S., Öztürk, H., & Tağa, T. (2014). Türkçe dersinde uygulanan sınavlar üzerine bir değerlendirme. *International Journal of Social Science*, 28, 513–523.
- YÖK (2018). Öğretmen yetiştirme lisans programları. <http://www.yok.gov.tr>
- Yüceer, D. (2024). Türkçe öğretmeni adaylarının soru hazırlamaya dair deneyim ve görüşleri. *Ana Dili Eğitimi Dergisi*, 12(1), 69–86.