# 10. BÖLÜM

# PUANLAMA YÖNTEMLERİ

***SCORING METHODS***

***Doç. Dr. Murat Doğan ŞAHİN***

# PUANLAMA YÖNTEMLERİ

***Doç. Dr. Murat Doğan ŞAHİN***
*Anadolu Üniversitesi*

**Öz**

Bu bölümde eğitimde kullanılan puanlama yöntemleri, puanın nasıl üretildiğini merkeze alan epistemolojik bir bakış açısıyla ele alınmıştır. Bu doğrultuda yöntemler; puanlama sürecine dayalı yaklaşımlar, otomatik puanlama ve karşılaştırma yargılarıyla puanlama olmak üzere üç temel başlık altında incelenmiştir. İlk bölümde manuel ve dijital puanlama, ikili ve kısmî puanlama ile ağırlıklı puanlama yaklaşımları; kullanım alanları, avantajları, sınırlılıkları ve psikometrik özellikleri bakımından değerlendirilmiştir. Ayrıca puanlama sürecinin ölçme amacına, ölçülen özelliğe ve uygulama koşullarına göre nasıl farklılaşabileceği tartışılmıştır. İkinci bölümde otomatik puanlama sistemlerinin tarihsel gelişimi, geleneksel kural ve özellik temelli modellerden derin öğrenme ve üretken yapay zekâ temelli yaklaşımlara uzanan dönüşüm çerçevesinde açıklanmıştır. Kural temelli, makine öğrenmesi, sinir ağı ve ontoloji temelli yöntemlerin çalışma ilkeleri özetlenmiş; CNN, LSTM ve Transformer mimarilerinin yazılı performansların değerlendirilmesindeki rolleri ele alınmıştır. Büyük dil modellerinin otomatik puanlamaya sağladığı olanaklar, zero-shot, few-shot ve fine-tuning yaklaşımları kapsamında incelenmiş; otomatik puanlama performansının değerlendirilmesinde *QWK* başta olmak üzere kullanılan ölçütler ve bu ölçütlerin yorumlanmasına ilişkin sınırlılıklar tartışılmıştır. Son bölümde ise karşılaştırma yargılarıyla puanlama yaklaşımı tanıtılmış; mutlak puanlama anlayışından farklı olarak performansların ikili karşılaştırmalar yoluyla değerlendirilmesine dayanan bu yöntemin kuramsal temelleri açıklanmıştır. Böylece bölüm, geleneksel insan puanlamasından yapay zekâ destekli sistemlere ve karşılaştırmalı yargı temelli yaklaşımlara kadar güncel puanlama yöntemlerini bütüncül bir bakış açısıyla değerlendirerek, farklı ölçme amaçları ve uygulama koşulları için uygun puanlama yaklaşımlarının seçimine yönelik kuramsal ve uygulamalı bir çerçeve sunmaktadır.

***Anahtar Kelimeler***

*Manuel ve Dijital Puanlama, Kısmî Puanlama, Ağırlıklı Puanlama, Otomatik Puanlama, Üretken Yapay Zekâ, Karşılaştırma Yargılarıyla Puanlama*

# SCORING METHODS

***Assoc. Prof. Dr. Murat Doğan ŞAHİN***
*Anadolu University*

**Abstract**

This chapter examines scoring methods used in educational assessment from an epistemological perspective that focuses on how scores are generated. Accordingly, the methods are organized into three main categories: process-based scoring approaches, automated scoring, and comparative judgement. The first section discusses manual and digital scoring, dichotomous and partial-credit scoring, and weighted scoring approaches in terms of their applications, advantages, limitations, and psychometric properties. It also examines how scoring procedures vary according to the purpose of assessment, the construct being measured, and the conditions under which the assessment is administered. The second section explains the historical development of automated scoring systems, tracing their evolution from traditional rule-based and feature-based models to deep learning and generative artificial intelligence-based approaches. It summarizes the operating principles of rule-based, machine learning, neural network, and ontology-based methods, and discusses the roles of CNN, LSTM, and Transformer architectures in the evaluation of written performances. The opportunities provided by large language models for automated scoring are examined within the context of zero-shot, few-shot, and fine-tuning approaches. In addition, the chapter discusses the metrics used to evaluate automated scoring performance —particularly the *QWK* — and the limitations associated with interpreting these metrics. The final section introduces the comparative judgement approach to scoring and explains its theoretical foundations as an alternative to absolute scoring, in which performances are evaluated through pairwise comparisons. Overall, the chapter provides a comprehensive perspective on contemporary scoring methods, ranging from traditional human scoring to AI-assisted systems and comparative judgement approaches. It offers both a theoretical and practical framework for selecting appropriate scoring methods for different assessment purposes and implementation contexts.

***Keywords***

*Manual and Digital Scoring, Partial-credit Scoring, Weighted Scoring, Automated Scoring, Generative Comparative Judgement*

## 1. GİRİŞ

Bir performansın puanlanmasında kullanılabilecek çeşitli yöntemler bulunmaktadır. Hangi yöntemin kullanılacağı ölçmeye konu olan özellik başta olmak üzere ölçme işleminin amacı, grubun büyüklüğü ve değerlendirme işlemini yapan kurumun-bireyin operasyonel kabiliyeti gibi bir dizi değişkene göre farklılaşabilmektedir. Bu bölümde, eğitimde ve ölçme değerlendirme alanında sık kullanılan puanlama yöntemleri incelenmiştir. Söz konusu yöntemleri farklı başlıklarda ele almak mümkündür. Bu bölümde puanlama süreçlerini tarihsel olarak sınıflamak (örneğin, geleneksel ve yenilikçi gibi) tercih edilmemiştir. Çünkü teknolojideki hızlı gelişime paralel olarak bugün yenilikçi görünen bir uygulamanın kısa sürede yerini başka uygulamalara bırakması ve birkaç yıl içinde geleneksel olarak ifade edilmesi muhtemeldir. Buna bağlı olarak mevcut bölümde puanlama yöntemleri "puanlama sürecine dayalı yaklaşımlar", "otomatik puanlama" ve "karşılaştırma yargılarıyla puanlama" olmak üzere üç kısımda sunulmuştur. Bu ayrım, puanlama sürecinin epistemolojik olarak ele alınmasıyla ortaya çıkmıştır. Puanlama yaklaşımlarının epistemolojisi merkeze alındığında yöntemler teknolojiye veya uygulama biçimine göre kategorize edilmez; puanın bir bilgi olarak nasıl üretildiği, hangi varsayımlara dayandığı ve performansa ne tür anlamlar yüklediği temele alınarak bir sınıflandırma yapılır.



Puanlama süreçleri, cevaplayıcı kararlarına dayalı olarak da yürütülebilir. Bu puanlama biçimi sınava katılanların kesinlikle yanlış olduğuna yönelik düşüncelerini ya da işaretledikleri seçeneğin doğruluğuna olan güvenlerini esas alır (Özdemir, 2003). Bu bölüm kapsamında yanıtlayıcıları merkeze alan yöntemlere yer verilmemiştir.

## 2. PUANLAMA SÜRECİNE DAYALI YAKLAŞIMLAR

Puanlama sürecine dayalı yaklaşımlar; puanın önceden belirlenmiş kurallara göre üretildiği, karar akışının açık, öngörülebilir ve şeffaf olduğu puanlama sistemlerini ifade eder. Bu yaklaşımda puan, algoritmik tahminlerden değil, tanımlı bir değerlendirme sürecinin sonucundan türetilir. Bu nedenle belirlenen puanlama yöntemi kâğıt üzerinde ya da dijital ortamda olması fark etmeksizin aynı biçimde uygulanır. Bir başka ifadeyle, puanlamanın yapıldığı ortamın yanı sıra kullanılan puanlama sürecinin niteliği ve puanın nasıl üretildiğine ilişkin ilkeler bu yaklaşımın sınıflandırılmasında belirleyici rol oynar. Bu doğrultuda puanlama sürecine dayalı yaklaşımlar puanın verildiği ortam, puanın veriliş biçimi olmak, otomatik puanlama ve karşılaştırma yargılarıyla puanlama olmak üzere dört kısımda incelenmiştir.

## 2.1. Puanlama Yapılan Ortama Göre Puanlama Türleri

Bu başlık altında, puanlamanın geleneksel metotlarla (kâğıt-kalem vb.) ya da dijital ortamlarda yapılmasını esas alan yöntemlere değinilmiştir.

### 2.1.1. Manuel Puanlama

Manuel puanlama, öğrenci yanıtlarının veya performanslarının doğrudan bir insan tarafından puanlanmasıdır. Bu yöntemde puanlayıcı, yanıtları inceleyerek önceden belirlenen puanlama ölçütlerine göre karar verir. Açık uçlu yanıtlar, üretimsel dil becerileri (yazma ve konuşma), sunumlar, yaratıcılık gerektiren performanslar ve uygulamalı performans görevleri gibi ürünlerin değerlendirilmesi insan puanlayıcılar tarafından yapılabilir. Özellikle yapılandırılmış eğitim ortamlarındaki puanlama işlemleri genellikle bu yolla yapılır.

Puanlama sürecine esneklik kazandırması manuel puanlamanın kayda değer bir üstünlüğüdür. Ancak bu yöntemin bazı sınırlılıkları da vardır. Öncelikle, süreç zaman alıcıdır ve büyük öğrenci gruplarında uygulanması oldukça zahmetlidir. Ayrıca, puanlamada subjektiflik ortaya çıkabilir; aynı performansın farklı puanlayıcılar tarafından farklı şekilde puanlanması, güvenirlik sorunlarına yol açabilir. Bu tür sorunları azaltmak için puanlama anahtarları, kontrol listeleri, dereceleme ölçekleri veya rubriklerden (dereceli puanlama anahtarlarından) yararlanılması önerilir. Fakat özellikle K-12 düzeyinde yapılan performans değerlendirmelerinde bir puanlama anahtarına bağlı kalmaktan ziyade genel izlenimle puanlama yaygındır. Bu da sonuçların adilliği ve objektifliği konusunda endişeleri beraberinde getirmektedir.

Manuel puanlamada yoğun bir insan emeği söz konusudur; puanlama sürecindeki tüm işlemler insan eliyle yapılır. Puanlanacak performans sayısının az olduğu, nispeten düşük önem arz eden uygulamalarda ve dijital araçlardan yararlanmak mümkün olmadığında manuel puanlama yapılması makul görülebilir. Manuel puanlama, teknolojik altyapının yetersiz olduğu durumlarda zorunlu bir seçenek olmasının yanında, özellikle üst düzey düşünme becerilerinin ölçülmesinde ve nitelikli geri bildirim sunulmasında hâlâ önemli bir yere sahiptir. Nitekim bölümün devamında üzerinde durulacak yapay zekâ uygulamalarının üst düzey becerilerinin değerlendirilmesinde kullanılabilmesi çok maliyetli olabilmekte ve kimi zaman bu maliyetlere rağmen beklenen sonuçlar alınamamaktadır.

### 2.1.2. Dijital Puanlama

Dijital değerlendirme veya elektronik değerlendirme; sınavların, testlerin, anketlerin ya da ölçme işlemlerinin bilgisayar, tablet ve cep telefonu gibi dijital aygıtlar aracılığıyla sunulmasıdır (Assessment Systems, 2025). Sınavların dijitalleştirilmesi, yenilikçi ölçme-değerlendirme tasarımlarını kullanmayı mümkün kılmaktadır (OECD Digital Education Outlook, 2023). Bu durum puanlama süreçlerine de yansımaktadır. Dijital puanlama, öğrenci performanslarının dijital araçlar, yazılımlar veya sistemler aracılığıyla puanlanmasını ifade eder. Bu yöntem, değerlendirme sürecinde insan puanlayıcıların rolünü azaltabilir, destekleyebilir ya da tamamen ortadan kaldırabilir. Dijital puanlama sistemleri, teknolojik altyapının gelişmesiyle birlikte özellikle büyük ölçekli sınavlarda, çevrim içi ölçme ortamlarında ve merkezî değerlendirme uygulamalarında yaygın olarak kullanılmaktadır.

Dijital puanlama, zaman ve maliyet açısından verimlilik sağlayan, çok sayıda verinin kısa sürede işlenmesine olanak tanıyan bir yaklaşım sunar. Bu bağlamda kullanışlılığı artıran bir yöntem olduğu söylenebilir. Bununla birlikte, sistemin doğruluğu ve güvenilirliği kullanılan yazılımın niteliğine, algoritmaların hassasiyetine ve sistemin tasarımına bağlıdır.

#### *2.1.2.1. Optik Okuyucu ile Puanlama*

Genellikle çoktan seçmeli testlerde kullanılan, yanıtların özel formlara işaretlendiği ve bu formların optik okuyucu cihazlar tarafından otomatik olarak değerlendirildiği bir dijital puanlama yöntemidir. Bu sistem, özellikle kalabalık gruplara yapılan uygulamalarda kısa sürede ve hatasız biçimde puanlamanın esas olduğu durumlar için ideal bir çözümdür. Bu da çoktan seçmeli maddeleri geniş ölçekli sınavlar için oldukça kullanışlı hale getirir.

Optik okuyucu sistemleri, işaretlenen cevapları belirli bir anahtarla karşılaştırarak doğru ve yanlışları belirler; böylece hızlı ve objektif bir puanlama süreci sağlar. İnsan hatasını en aza indirir, değerlendirme sürecini büyük ölçüde otomatikleştirir ve raporlama işlemlerini kolaylaştırır. Bununla beraber optik okuyucu sistemlerinin sınırlı bir değerlendirme kapasitesi vardır. Sadece seçme gerektiren maddeleri (Bkz. 6. Bölüm) işleyebildiği için açık uçlu, yorum gerektiren, yaratıcı ya da analitik düşünme becerilerini ölçmeye yönelik yanıtları puanlayamaz. Sonuç olarak, optik okuyucu ile puanlama; hız, doğruluk ve maliyet açısından avantajlı bir sistem sunmakla birlikte madde türü çeşitliliği bakımından sınırlayıcıdır. Bundan dolayı, bu yöntemin uygunluğu ölçülmek istenen beceri düzeyi ve sınavın amacı doğrultusunda dikkatle ele alınmalıdır.

#### *2.1.2.2. Performansın Dijital Ortamda Puanlanması*

Bazen ölçme işlemine konu olan özellikleri çoktan seçmeli maddelerle ölçmek mümkün olmaz. Örneğin, yazma ya da konuşma gibi performans gerektiren özellikler çoktan seçmeli maddeler ile ölçülemez. Söz konusu özellikler performans görevleri aracılığıyla ölçülürken sınavın tüm süreçleri ya da sadece değerlendirme basamağı dijital ortamda gerçekleştirilebilir. Özellikle bilgisayar tabanlı sınavlarda öğrencinin yanıtı dijital ortamda sözlü ya da yazılı olarak (ölçülen özelliğe göre) alınır; böylece değerlendirme süreci de çoğunlukla rubriklere dayalı olarak bu dijital ortamda ilerler.

Kimi zaman ise sınav kâğıt-kalem biçiminde yapılır ancak yanıtlar dijital ortama aktarılarak sadece puanlama süreci bu ortamda gerçekleşir. İster tüm süreç isterse yalnızca puanlama aşaması dijital ortamda yürütülsün performansın belirli düzeyleri olduğunda kısmî puanlamalar yapılabilir. Burada, değerlendirme otomasyonunun özellikleri ön plana çıkar. Buna göre sınavın önem düzeyi dikkate alınarak her bir yanıt sadece bir ya da birden çok puanlayıcı tarafından puanlanabilir. Yüksek risk taşıyan sınavlarda puanlamalar genelde birden fazla puanlayıcı tarafından gerçekleştirilir. Puanlayıcı sayısının birden fazla olduğu bu tür durumlarda, puanlayıcıların birbirinden bağımsız ve bir rubriğe dayalı olarak yaptıkları puanlamalar arasındaki uyuma göre süreç devam eder ve performansa nihai puan verilir. Sınıf içi ölçme-değerlendirme süreçlerinde ise çoğu zaman sadece bir puanlayıcı performansa değer biçer.

Özellikle bilgisayar tabanlı sınavlarda, madde türü çoktan seçmeli ya da doğru-yanlış türünde olduğunda optik okuyucuya gerek kalmaz; öğrenci yanıtladığı anda yanıtın doğru/yanlış olduğu değerlendirilebilir. Bu özellikleriyle dijital ortamlar çoktan seçmeli testlerin bireye uyarlanmış biçimde yapılabilmesini mümkün kılar.

## 2.2. Puanın Veriliş Biçimine (Puanlama Kuralına) Göre Puanlama Türleri

Puanlama sürecinde işletilen kural ya da başka bir ifadeyle puanın veriliş biçimi, kullanılan ölçme aracının özelliğiyle ilgilidir. Ölçme aracında çoktan seçmeli, doğru-yanlış ya da açık uçlu madde türlerinden hangisine yer verildiği puanlama biçimini etkiler.

### 2.2.1. İki Kategorili Puanlama (1-0 Puanlama)

Kısmi puanlamanın kullanılmadığı durumlarda maddeye verilen yanıt ya tam doğru ya da tamamen yanlış kabul edilir. Bu tür durumlarda elde edilen veri türü "1-0" olarak anılır. Yanıt doğruysa 1, yanlışsa veya yanıt yoksa 0 puan verilir (Kurz, 1999). Doğru-yanlış türündeki maddeler 1-0 şeklinde puanlanır. Bunun yanında çoktan seçmeli ve kısa cevaplı maddelerde de çoğunlukla 1-0 şeklinde puanlama yapılır. Maddelerin bu şekilde puanlandığı ölçme araçlarında toplam puanı hesaplamak için doğru yanıt sayısı her doğru yanıt için ön görülen katsayıyla çarpılır. Şans başarısını azaltmak amacıyla düzeltme işlemi uygulandığında ise seçenek sayısına bağlı olarak her yanlış yanıt, doğru

sayının belirli bir yüzdede düşmesine neden olur. Böylece "net", yani düzeltilmiş puan hesaplanır ve sonrasında belirlenen katsayıyla çarpılır. 1-0 şeklinde puanlanan maddeler genelde aynı ağırlığa sahiptir; daha açık bir anlatımla her maddenin nihai test puanına katkısı eşittir.

### 2.2.2. Kısmi Puanlama (Çok Kategorili Puanlama)

Özellikle geniş ölçekli uygulamalarda kullanışlılık ilkesi gereği çoğunlukla 1-0 şeklinde puanlanan maddelere başvurulur. Bu tür puanlama, ölçülen özelliğin bireyde ya tam olarak var olduğunu ya da hiç olmadığını, bir başka ifadeyle bireyin ya başarılı ya da başarısız olduğunu kabul eder. Ancak çoğu zaman soruyla yoklanmak istenen özellik, bireylerde farklı düzeylerde görülebilir. Örneğin, yazılı ya da sözlü anlatım becerisinin ölçülmesi hedefleniyorsa 1-0 şeklinde puanlanan bir madde türü kullanmak geçerlik problemi yaratır. Dolayısıyla bu tür beceriler ölçülürken performans görevlerinin kullanılması ve kısmi puanlama yapılması tercih edilir.

Öğrenci performansının farklı düzeylerde puanlanmasına kısmî puanlama adı verilir. Önceden hazırlanan bir cevap anahtarı ya da dereceli puanlama anahtarına dayanarak bireylerin yanıtları farklı düzeylerde puanlanabilir. Kısmi puanlamanın düzeyleri, ölçme işleminin amacına ve ölçülen özelliğe göre farklılık gösterebilir. Kimi zaman bir performans "yanlış-geçersiz yanıt, kısmen doğru, tam doğru" olmak üzere üç seviyeye ayrılır. Üst düzey becerilerin ölçüldüğü durumlarda ise düzey sayısı artabilir. Kısmi puanlamada düzey sayısı ölçülen özellik kadar, sınavın amacına ve eldeki imkânlara da bağlıdır. Sınıf içinde yapılan ölçme-değerlendirme işlemlerinde daha esnek bir puanlama yaklaşımı mümkünken geniş ölçekli uygulamalarda, özellikle yüksek riskli sınavlarda, puanlama güvenirliğini yüksek tutmak amacıyla düzey sayısı daha katı biçimde belirlenir. Bu noktada duyarlılık ile birbirinden net biçimde ayrılabilecek düzey sayısı arasında bir denge gözetmek gerekir.

> Anahtarla puanlamada bütünsel ve analitik yaklaşım tercih edilebilir. Ölçme işlemine konu olan özellik bütünsel değerlendirmeye uygunsa holistik (bütünsel) bir puanlama aracı kullanılır. Öte yandan özellik aşamalık arz ediyor ya da alt bileşenlerden oluşuyorsa bunlardan her birini ayrı ayrı puanlamak gerekir ki bu durumda analitik bir puanlama aracı tercih edilir (Kula-Kartal, 2021).

Kısmi puanlama, çoktan seçmeli maddelerde de kullanılabilir; buna göre tam doğru seçeneğin yanında kısmi puanlama yapılacak seçenekler de belirlenebilir. Doğru yanıtı verene kadar yanıtlama, tamamen yanlış bilgiden başlayarak çeşitli düzeylerde kısmi ve tam bilgiye ulaşmayı esas alan eleme puanlaması ve adaylara seçeneklerin herhangi bir sayıda doğru olabileceğinin bildirildiği çoklu seçim formatı (Bkz. 6. Bölüm), çoktan seçmeli maddelerde kullanılması mümkün olan kısmi puanlama yöntemleridir (Kurz, 1999). Uygulanmasındaki karmaşıklığa karşın güvenirlik ve geçerlik açısından getirilerinin sınırlı olması bu metotların yaygınlaşmasına engel teşkil etmektedir.

### 2.2.3. Ağırlıklı Puanlama

Klasik Test Kuramı'na (KTK) dayalı uygulamalarda çoktan seçmeli maddeler, yer aldığı alt testte genel olarak eşit ağırlığa sahiptir. Fakat bazen testteki maddelerden bazılarının diğerlerine göre daha yüksek ya da daha düşük ağırlığa sahip olması gerekli görülür. Bu gereklilik, çoktan seçmeli testlerin 1-0 şeklinde puanlanmasının getirdiği sınırlılıklardan kaynaklanır. Abu-Sayf (1979), dört temel sınırlılıktan bahsetmektedir. Bunlardan ilki, testin yönergelerine dayanarak tahmini teşvik etmenin görece yüksek oranda hata varyansı oluşturduğu ve bu formülün kısmi bilgiyi hesaba katmadığı yönündeki psikometrik argümandır. Pragmatik argüman, temkinli davranarak bilmediği soruları boş bırakan adayların risk alanlara kıyasla cezalandırıldığı yönündedir. Ahlaki argüman tahmin yürütmenin yanlış ve bu davranışı ödüllendirmenin sakıncalı olduğu görüşüne dayanırken siyasi argüman adayları tahminde bulunmaya teşvik etmenin çoktan seçmeli testlere olan güvenlerini kaybetmelerine yol açacağını iddia eder. Bunun yanında kimi maddelerin diğerlerinden daha kritik becerileri ölçmesi, daha zor veya ayırt edici olması ya da ikili puanlamanın yanlış yapan tüm yanıtlayıcıları ölçülen beceriden aynı düzeyde yoksun sayması (Rowley & Traub, 1977) gibi sınırlılıklar çoktan seçmeli maddelerde ağırlıklandırmayla puanlamaya yönelik çalışmalara zemin hazırlamıştır. 1900'lü yılların başlarından itibaren bu konudaki araştırmalar hız kazanmıştır. Bununla birlikte Stanley ve Wang (1970) o güne değin yapılan çalışmaları incelediğinde ağırlıklandırmanın test puanlarının güvenirlik ve geçerliğinde kayda değer bir iyileşme yaratmadığını ortaya koymuştur. Ağırlıklı puanlama genellikle önsel ve sonsal olmak üzere iki başlıkta ele alınır.

#### *2.2.3.1. Önsel Ağırlıklı Puanlama*

Kimi zaman uzmanlar, testte yer alan maddelerin bir kısmının daha kritik becerileri ölçtüğü ya da ilgili maddeleri yanıtlayabilmek için daha uzun süreli/üst düzey bir performans ortaya konulması gerektiği görüşünde olurlar. Ya da test sadece çoktan seçmeli maddelerden oluşmuyor, açık uçlu maddeler de içeriyor olabilir. Böyle durumlarda, uygulama öncesinde uzman görüşlerine dayanarak maddelerin nihai puanlamadaki ağırlıkları belirlenebilir ve kararlaştırılan bu ağırlıklar doğrultusunda her birey için nihai puan hesaplanır. Ancak bu yöntemde belirlenen ağırlıkların keyfî olmadığından emin olunmalıdır. Bunun için tek bir alan uzmanının görüşünden hareketle karar vermekten kaçınılmalı, birden fazla uzmanın görüşüne başvurulmalıdır.

Önsel olarak özellikle çok kategorili puanlama ile ağırlıklandırılan maddelerin güvenirliğe etkisi tartışma konusudur. Benton (2018) yaptığı geniş çaplı araştırmada, çok kategorili puanlanan maddelerin aynı toplam puan değerinde ancak bireysel olarak puanlanmış maddelere kıyasla (söz gelimi, 0-4 arası puanlanabilen tek bir madde ile her biri 1-0 şeklinde puanlanan dört madde) güvenirliğe katkısının daha düşük olduğunu belirlemiştir. Düşük puanlı maddelerin ağırlığı artırılıp çok kategorili olanların

azaltılmasıyla elde edilen puanların ilk puanlarla korelasyonun çok yüksek olduğu, düzeltme sonucunda ulaşılan çıktıların böyle bir işlemi yapmaya gerekçe oluşturmayacağı ilgili çalışmada rapor edilen dikkat çekici bir diğer bulgudur.

#### *2.2.3.2. Sonsal Ağırlıklı Puanlama*

Testteki maddelerin ağırlıklandırılmasına ilişkin önsel güçlü bir dayanak olmadığında ya da bu tür bir dayanağın keyfiyet arz edeceği düşünülüp ampirik kanıtlara gereksinim duyulduğunda ağırlıklandırma için testin uygulanmasıyla elde edilen istatistiklerden yararlanılabilir. Sonsal ağırlıklandırma denilen bu yöntemde, maddeler testin uygulanması sonucunda ulaşılan istatistiklere dayalı olarak ağırlıklandırılır. Burada akla ilk gelen yöntem, maddenin ayırt edicilik değeri ile puanın çarpılmasıdır. Böyle bir durumda madde 1-0 şeklinde puanlanmaz. Yanlış yanıta 0 puan verilirken doğru yanıta 0-1 arasında kalan ve değeri maddenin ayrıt edicilik indeksine (*r*) eşit olan bir puan atanır. Birnbaum (1968) çalışmasında bireylerin test puanlarını, doğru yanıtlarını madde ayıt ediciliğiyle çarparak elde etmiştir. Bir başka yöntem ise maddenin güçlüğünü esas almaktır. Bu durumda madde güçlük indeksi (*p*) arttıkça, yani madde kolaylaştıkça bu maddeden alınan puan düşer; tersine madde zorlaştıkça madde puanı artar. Ancak bir maddenin güçlük değeri onun işlerliğine ilişkin tek başına bilgi vermediğinden bu yöntemin kullanımı konusunda çekinceler vardır. Günümüzde hala yaygın olarak başvurulan bir başka yöntem şans başarısının düzeltme amacıyla kullanımına dayanır. Bu yöntemde bireyin yanlış sayısı testteki maddelerin seçenek sayısının bir eksiğine (örneğin, 5 seçenekli bir madde için 4) bölünür ve elde edilen değer doğru sayısından çıkarılarak testte puanlamaya esas teşkil eden "net" belirlenir. Kelley (1947) ise sonsal ağırlıklandırma için testten alınan puanlara ait güvenirlik kestiriminin kullanıldığı bir regresyon modeli önermiştir (akt: Yurdugül, 2010).

Benton (2018), maddeleri ağırlıklandırırken güvenirliğin maksimize edilmesini esas alan bir yöntem kullanmıştır. Elde ettiği sonuçlar, güvenirlikteki yükselmenin testin yordama geçerliği açısından çoğunlukla bir artışla sonuçlanmadığını göstermiştir. Kılıç ve Doğan (2019) ise yeni bir ağırlıklandırma yöntemi önermiştir. Buna göre bireyin testteki doğru sayısı testte yer alan madde sayısına bölünür ve bu bölme işleminden elde edilen değere maddenin güçlüğü eklenir. Hesaplanan nihai değer 1'in altında çıkarsa bireyin aldığı puan 1 ya da 0 olarak aynen devam eder, herhangi bir işlem yapılmaz. Nihai değer 1'e eşit ya da 1'den büyükse bireyin ilgili maddeyi doğru yanıtlayıp yanıtlamadığına göre hareket edilir; birey maddeye doğru yanıt vermişse 1'in üzerine maddenin güvenirlik değeri eklenirken yanlış yanıt vermişse herhangi bir işlem yapılmaz. Söz gelimi bir maddenin güçlük değeri 0,36, ayırt edicilik indeksi 0,56 ve bireyin testteki ortalaması (doğru yanıtlarının sayısı/testteki madde sayısı) 0,70 olsun. Bu durumda Güçlük + Bireyin test ortalaması (0,36 + 0,7 = 1,06) 1'in üzerinde çıkar. Dolayısıyla bireyin doğru yaptığı bu madde için 1'in üzerine madde güvenirliği eklenir ve bireyin maddeden aldığı puan 1+0,27=1,27 olur. Araştırmacılar, bu yöntemin testin güvenirliğine ve açıklanan varyans oranına katkıda bulunduğu sonucuna ulaşmışlardır.

Yaşar ve arkadaşları (2021) kullandıkları yöntemde maddelerin güçlük indeksine dayalı bir ağırlıklandırma yapmayı denemişler ve uyguladıkları ağırlıklandırmanın testin aritmetik ortalamasında düşüş, standart sapmasında ise artış ile sonuçlandığını rapor etmişlerdir. Diğer bir deyişle bu yöntem uygulandığında testin bireyler arasındaki farklılıkları ortaya koymada daha başarılı olduğunu ortaya koymuşlardır.

KTK uygulamalarında ağırlıklandırma, sınıf içi uygulamalar ya da düşük önem arz eden sınavlar dışında pek tercih edilen bir yöntem değildir. Bir maddenin güçlük ya da ayırt edicilik indeksine göre puanlamaya etkisinin farklı olması gerektiği düşünülüyorsa bir diğer ölçme kuramı olan Madde Tepki Kuramı'ndan (MTK) yararlanmak gerekir. Sonsal ağırlıklandırma yapılan birçok çalışmada MTK temelli uygulamalara da başvurulmaktadır. Bununla birlikte bu konuya ilişkin detaylı bilgilere kitabın 2. bölümünde yer verilmiştir.

## 2.3. Otomatik Puanlama

Otomatik puanlama, bir performansın insan puanlayıcıların değerlendirmelerine benzer biçimde analiz edilmesini ve puanlanmasını sağlayan, yapay zekâ veya istatistiksel modeller kullanan bilgisayar temelli bir sistemdir. Bu sistemler, özellikle doğal dil işleme ve makine öğrenmesi tekniklerinden yararlanarak yanıtların dilsel, anlamsal ve yapısal özelliklerine göre değerlendirilmesini esas alır. Bu yolla açık uçlu maddelere verilen yanıtların daha hızlı ve ekonomik biçimde puanlanması mümkün olabilmektedir. Otomatik puanlama sonucu elde edilen puanların geçerliğinin sağlanması bu süreçteki en önemli husustur. Zira birçok algoritmayla otomatik değerlendirme yapılarak bir puan elde edilebilmektedir. Bu puanın öğrencinin gerçek performansını ortaya koyması otomatik puanlama algoritmalarının işlerliği için en önemli ölçüttür. Kullanılan algoritmanın puanlama sırasında metindeki hangi özellikleri referans alacağı ya da farklı özellikleri aynı anda kullanırken nasıl bir ağırlıklandırma yapacağı algoritmanın performansını etkilemektedir.

Temeli 1960'lara dayanmakla beraber otomatik puanlamadaki önemli gelişmeler ve araştırmalar yaklaşık 20 yıllık bir geçmişe sahiptir. Otomatik puanlamanın erken örneklerinde genellikle dilbilgisi, sözcük dağarcığının karmaşıklığı ve cümle yapısı gibi iyi tanımlanmış sınırlı sayıdaki dilsel özellik puanlamada kullanılmıştır. Bu özellikler, ölçülmek istenen yapıyı yakından temsil edecek ve kapsayacak biçimde alan uzmanları tarafından titizlikle seçilir. Bu nedenle elde edilen puanlar yüksek düzeyde açıklanabilir ve yorumlanabilir niteliktedir. Bu erken dönem yaklaşımlarda modelde kullanılan özellikler ölçülmek istenen yapıyla doğrudan ilişkili olup her bir özelliğe atanan ağırlıklar açık biçimde görülebilmekte ve modelin bu özelliklerden hareketle nasıl bir puanlama süreci izlediği adım adım takip edilebilmektedir. Dolayısıyla bu tür modellerin ürettiği puanlar yüksek derecede şeffaf, açıklanabilir ve yorumlanabilir nitelik taşımaktadır (Zhang ve ark., 2024). Fakat bir metindeki anlam örüntüsünden ziyade

metinsel özelliklere yöneldiğinden (uzunluk, belirli kelimelerin kullanımı vb.) bu modeller kimi zaman güçlü anlamsal içeriklere hak ettiğinin altında bir değer biçebilmektedir (Aydın ve ark., 2025). Ayrıca geleneksel modeller, yanıtların bir bölümünün önceden insan puanlayıcılar tarafından puanlanmasını zorunlu kılar ve bu verilere dayanarak insan puanlayıcıların kararlarını taklit edebilecek örüntüleri öğrenir.

Otomatik puanlama üzerine yapılan çalışmaların Page (1966) tarafından geliştirilen ve bir kompozisyonu dilbilgisi, sözcük seçimi ve cümle yapısı gibi özelliklere göre değerlendiren Project Essay Grader (PEG) ile başladığı söylenebilir. Foltz ve arkadaşları (1999) ise Intelligent Essay Assessor (IEA) adlı sistemi kullanarak örtük anlamsal çözümleme (latent semantic analysis) yöntemiyle içerik metnin değerlendirmesini yapmıştır. Shermis ve arkadaşları (2001) tarafından yayımlanan PEG'in geliştirilmiş bir versiyonu, dilbilgisine odaklanmış ve sistemin puanlama performansını insan puanlayıcılar ile karşılaştırmıştır. Powers ve arkadaşları (2002), E-rater sistemini; Rudner ve arkadaşları (2006), Intellimetric'i; Rudner ve Liang (2002) ise Bayesian Essay Test Scoring System'i (BETSY) geliştirmiştir. Bu sistemler, bir yazma performansını (bu örneklerde kompozisyonu) puanlamak için biçem (style) ve içerik özelliklerine odaklanan doğal dil işleme (NLP) tekniklerini kullanmaktadır (akt: Ramesh & Sanapudi, 2022).

1990'lı yıllarda geliştirilen otomatik puanlama sistemlerinin büyük çoğunluğu, örüntü eşleme (pattern matching) ve istatistik temelli yaklaşımları izlemiştir. Sonraki yıllarda kullanılan puanlama sistemleri regresyon temelli ve doğal dil işleme tekniklerini işe koşmaya başlamıştır. 2014 yılından itibaren geliştirilen otomatik puanlama sistemleri, sözdizimsel (syntactic) ve anlamsal (semantic) özellikleri içeren derin öğrenme tekniklerini kullanarak önceki modellere kıyasla daha başarılı sonuçlar elde etmiştir (Dong ve ark., 2017).

Geniş dil modellerindeki-GDM (Large Language Models-LLM) gelişmelerle birlikte daha önce kural ya da özellik temelli olarak iki kısımda ele alınan geleneksel yöntemlerin yerini hızla üretken yapay zekâ (ÜYZ) temelli otomatik puanlama almaya başlamıştır (Aydın ve ark., 2025; Bui & Barrot, 2025; Garcia-Varela ve ark., 2025; Lee ve ark., 2024). ÜYZ temelli otomatik puanlama uygulamalarıyla birlikte, geleneksel uygulamalardaki metodolojik ve dil bilimsel yetkinliklere sahip olma zorunluluğu önemli ölçüde azalmış, böylece otomatik puanlama çalışmalarında ve uygulamalarında kayda değer bir artış gözlenmiştir. Otomatik puanlama bölümünün bundan sonraki kısmında, buraya kadar ana hatlarıyla açıklanan yöntemler farklı başlıklarda derinlemesine ele alınmıştır.

### 2.3.1. Geleneksel Otomatik Puanlama Yöntemleri

Bu başlık altında ÜYZ temelli uygulamaların öncesinde kullanılan otomatik puanlama yaklaşımlarına yer verilmiştir. Bu yöntemleri kural temelli ve özellik temelli yöntemler şeklinde iki kısımda ele almak mümkündür.

Kural Temelli Yöntemler, metindeki anlam örüntüsünden ziyade bu anlam örüntüsüyle ilişkili olduğu düşünülen sayılabilir metinsel özelliklere odaklanır. Bu yaklaşımın altında, "yazının niteliği ölçülebilir özelliklerle temsil edilebilir" varsayımı yatar. Değerlendiricilerin yazılı performanslarını incelerken dikkate aldıkları "trins" (insan puanlayıcıların sezgisel olarak aradığı temel yazma özellikleri) olarak adlandırılan içsel özelliklerin belirlenmesini içerir. Ancak trins doğrudan gözlemlenemediğinden bunları temsil eden ölçülebilir göstergeler "prox"ların bilgisayar programları aracılığıyla hesaplanmasına çalışılır (Dikli, 2006). Örneğin, bir yazıdaki toplam kelime sayısı, "akıcılık" (fluency) trini için bir prox (yani dolaylı gösterge) kabul edilir. Erken dönem otomatik puanlama sistemlerinde yüzeysel özelliklere yapılan bu vurgu, içerik açısından metnin niteliğini artırmadan yalnızca biçimsel değişikliklerle (örneğin noktalama işaretleri ekleyerek) puan yükseltmeye yönelik hile stratejilerine sistemi açık hale getirmiştir. Bu sınırlılığı ortadan kaldırmak için sonraki yıllarda, yazma performansını puanlarken daha yordayıcı dilsel özellikler üzerinde duran başka doğal dil işleme araçları geliştirilmiştir. Söz gelimi Coh-Metrix, sözcüksel incelik (lexical sophistication), sözdizimsel karmaşıklık ve metin bağdaşıklığını (cohesion) inceleyen kapsamlı göstergelerden biridir (Yeung, 2025). Kısacası kural temelli yöntemlerde uzmanlar tarafından yazma performansının nesnel göstergeleri olarak kabul edilen metinsel özellikler önceden belirlenir ve puanlama işlemi buna göre yapılır; ileri düzey istatistiksel modellerden yararlanılmaz. Kural temelli yaklaşımlarda alan uzmanlarınca belirlenen içerik özelliklerinin puanlamada kullanılması, puanlamanın şeffaflığını artırdığı biçiminde yorumlanır.

Özellik temelli yöntemler ve makine öğrenmesi temelli yöntemleri aynı çatı altınca incelemek makul olacaktır. Çünkü özellik temelli yaklaşımlar makine öğrenmesine dayanan yaklaşımların erken biçimi olarak nitelendirilebilir. Bu yaklaşımda modellerin puanlanmış yanıtlarla eğitilmesi gerekir. Model, insanlar tarafından puanlanmış eğitim verisine dayalı olarak algoritmalar oluşturur ve bu algoritmaları kullanarak kalan yanıtları puanlar. Makine öğrenmesi temelli yöntemler kural temelli yöntemlerdeki özelliklerden bazılarını kullanabilir. Ancak bu durum önceden belirlenmemiştir; modelin eğitim sürecinde kullandığı ve insan puanlayıcıların verdiği puanları yordama düzeyi yüksek metinsel özellikler puanlamada rol oynayabilir. Makine öğrenmesi temelli yöntemler (1) *regresyon temelli*, (2) *sınıflama temelli* ve (3) *sinir ağı temelli modeller* ile (4) *ontolojik temelli yaklaşımlar* olmak üzere dört ana başlıkta sıralanabilir (Ramesh & Sanampudi, 2022):

*Regresyon Temelli Modeller*, en temel tanımıyla bağımlı/yordanan değişkenin en az aralık düzeyinde olduğu, yani yazılı performansa verilen puanların aralık düzeyinde kabul edildiği durumlarda kullanılır. Temelde bu istatistiksel metodoloji esas olmakla birlikte modelin performansını artırmak amacıyla farklı uygulamalardan yararlanılır. Metinler arası anlamsal benzerliği ortaya koymak amacıyla metin ve derlem (corpus) temelli

uygulamalar (Mohler & Mihalcea, 2009), metindeki farklı özelliklerin (features) her yazılı performans için vektörize edilmesine dayanan uygulamalar (Adamson ve ark., 2014), Ridge regresyon temelli uygulamalar (Sultan ve ark., 2014), bulanık ontoloji temelli (Darwish & Mohamed, 2020) ve metin madenciliği temelli uygulamalar (Süzen ve ark., 2020) regresyon temelli yaklaşımlara örnek olarak gösterilebilir.

*Sınıflama Temelli Modeller*, bağımlı/yordanan değişkenin aralık düzeyinin altında olduğu durumlarda kullanılır. Destek vektör makinelerinden yararlanılan yanıt temelli ve word2vec kullanılarak kosinüs benzerliğine odaklanan referans temelli uygulamalar (Sakaguchi ve ark., 2015), rassal orman temelli uygulamalar (Mathias & Bhattacharyya, 2018) ve *XGboost* temelli uygulamalar (Ke ve ark., 2019), bu yaklaşımlara örnek olarak gösterilebilir.

*Sinir Ağı Temelli Modeller* kısımda ise özellikle Evrişimsel Sinir Ağları – (Convolutional Neural Networks - CNN), Kısa-Uzun Süreli Bellek – (Long-short Term Memory- LSTM) ve hatta transformer temelli BERT gibi uygulamalar ön plana çıkar. Basit ileri beslemeli ya da yüzeysel sinir ağları geleneksel olarak makine öğrenmesi altında sınıflandırılmıştır. Ancak CNN ve LSTM ağları gibi derin öğrenme mimarilerinin ortaya çıkmasıyla bu modeller geleneksel makine öğrenmesi sınırlarının ötesine geçmiş ve giderek artan biçimde yapay zekâ temelli otomatik puanlama sistemleri bağlamında ele alınmaya başlanmıştır.

Ontoloji Temelli Yaklaşımlar**,** yazılı performansları anlamsal veya kavramsal ağlar biçiminde temsil eden, yüzeysel metin özelliklerinin ötesine geçerek kavramlar arası anlam ilişkilerini değerlendiren modellerdir. Sözcükler ve ifadeler alan-temelli ontolojilerle eşlenerek, metnin kavramsal tutarlılığı ve anlam ilişkileri değerlendirilir. Destek vektör regresyonu, rassal orman ve cümle benzerliği ağları gibi graf[1] temelli tekniklerle anlamsal benzerlik sayısallaştırılmıştır (Mohler ve ark., 2011; Ramachandran ve ark., 2015; Zupanc ve ark., 2017; Ajetunmobi & Daramola, 2017). Geleneksel yöntemlerde kullanılan metinsel özellikler Tablo 1'de sunulmuştur.

[1] Graf temelli yaklaşımlar (graph-based approaches), metni kavramlar (düğümler-nodes) ve bu kavramlar arasındaki anlamsal ilişkiler (kenarlar-edges) biçiminde temsil eden yöntemlerdir. Bu modeller, öğrencinin yazısındaki kavramsal yapı ve bağlantı örüntülerini analiz ederek yanıtın konuya ve beklenen bilgi yapısına (ontolojiye) ne ölçüde uygun olduğunu değerlendirir.

**Tablo 1**

*Özellik Türleri [Ramesh ve Sanampudi'den (2022) uyarlanmıştır]*

| İstatistiksel Özellikler | Biçem (Stil) Tabanlı Özellikler | İçerik Tabanlı Özellikler |
|---|---|---|
| Kelime sayısına göre metnin uzunluğu | Cümle yapısı | Belge içindeki cümleler arası bağdaşıklık |
| Cümle sayısına göre metnin uzunluğu | Sözcük türleri | Cümlelerin örtüşmesi (prompt ile) |
| Ortalama cümle uzunluğu | Noktalama | Bilginin uygunluğu |
| Ortalama kelime uzunluğu | Dilbilgisel yapı | Kelimelerin anlamsal rolleri |
| N-gram | Mantıksal bağlaçlar<br>Sözcük dağarcığı | Doğruluk<br>Tutarlılık<br>Temel kavramları ifade eden cümle |

Tablo 1'de özetlendiği üzere otomatik puanlama sistemlerinde kullanılan özellikler genellikle istatistiksel, biçemsel ve içerik temelli olarak sınıflandırılabilir. Hangi özelliklerin kullanıldığının yanı sıra bu özelliklerin bilgisayar ortamında nasıl temsil edildiği bir başka deyişle nasıl kullanıldığı da önemlidir. *Bag-of-Words* ve TF-IDF gibi sıklık temelli modellerden, anlamsal temsiller (embeddings) sağlayan derin öğrenme yaklaşımlarına kadar farklı metin temsil yöntemleri, modelin dili nasıl "gördüğünü" belirler. Bu yöntemler Tablo 2'de özetlenmiştir.

**Tablo 2**

*Metin Temsil Yöntemleri*

| Yöntem | Tanım | N-gram ile İlişkisi | Otomatik Puanlamada Kullanımı |
|---|---|---|---|
| **Bag of Words (BoW)** | Metni, kelimelerin sıklıklarına göre temsil eden, kelime sırasını dikkate almayan yöntemdir (Jurafsky & Martin, 2023). | N=1 olan özel bir durum olarak görülebilir. N-gram modelinin temelini oluşturur. | Yazılardaki kelime çeşitliliğini ve sıklık örüntülerini analiz etmek için erken otomatik puanlama sistemlerinde kullanılmıştır. |
| **TF-IDF (Term Frequency – Inverse Document Frequency)** | Kelimelerin bir belgede ne kadar önemli olduğunu, diğer belgelerdeki sıklığına göre ağırlıklandıran yöntemdir (Manning ve ark., 2008). | BoW temsiline dayanır, ancak kelimelere anlamlı ağırlıklar atar. | Önemli kelimeleri vurgulamak ve yüzeysel metin özelliklerinin ötesine geçmek için kullanılır. |
| **N-gram*** | Metindeki ardışık n kelimeden oluşan dizileri temel alan temsil yöntemidir (Shermis & Burstein, 2013). | BoW'un genişletilmiş hâlidir; kelime sırasını ve yerel bağlamı kısmen yakalar. | Dilsel akış, sözdizimi ve bağdaşıklık özelliklerini yakalamak için kullanılır. |
| **Word Embeddings (Word2Vec, GloVe, BERT)** | Kelimeleri çok boyutlu anlamsal uzaylarda vektör olarak temsil eden yöntemlerdir (Mikolov ve ark., 2013). | N-gram'ın ötesine geçerek anlamsal (semantik) ilişkileri doğrudan öğrenir. | Derin öğrenme temelli modellerde metnin anlamını yakalamak için kullanılır. |

* N-gram'lar hem dilsel bir özellik hem de bir metin temsil yöntemi olarak ele alınabilir. Özellik türleri bağlamında ölçülebilir bir dilsel örüntü, temsil yöntemleri bağlamında ise bu örüntülerin sayısal olarak kodlanma biçimi olarak değerlendirilmiştir.

Bu yöntemler, otomatik puanlamanın ilk dönemlerinde özellik mühendisliğine dayalı sistemlerin temelini oluşturmuştur. Ancak derin öğrenme mimarilerinin gelişmesiyle birlikte, bu özellikleri elle tanımlama gerekliliği ortadan kalkmış, modeller anlamı doğrudan öğrenmeye başlamıştır. Sıradaki başlıkta, bu dönüşümün dayandığı yapay zekâ temelli yaklaşımlardan bahsedilmiştir.

### 2.3.2. Yapay Zekâ Destekli Otomatik Puanlama

Geleneksel otomatik puanlama yöntemleri açıkça belirtilmiş kurallara ya da özellik mühendisliğine odaklanan yordayıcı yaklaşımlara dayanırken yapay zekâ destekli yaklaşımlar metnin içindeki anlam örüntüsünü ortaya koymaya yönelik ileri düzey matematiksel modeller üzerine kuruludur. Makine öğrenmesi temelli modellerde de bahsi geçen derin öğrenme yaklaşımları, günümüzde kullanılan yapay zekâ destekli otomatik puanlama yaklaşımlarının öncül adımları olarak nitelendirilebilir.

#### *2.3.2.1. Derin Öğrenme Temelli Otomatik Puanlama*

Bu yöntemlerden ilk akla gelen CNN ve LSTM'dir. CNN, yerel veya ardışık girdilerdeki "öğrenilebilir filtreler" (convolutions) aracılığıyla örüntüleri yakalayan derin öğrenme mimarisidir. LSTM ise ardışık verilerde uzun süreli bağımlılıkları/örüntüleri yakalamaya yönelik bir RNN (tekrarlayan sinir ağı-recurrent neural nets) yapısı olup geleneksel RNN'lerin "uzun bağlamı öğrenememe" sorununu gidermek üzere geliştirilmiştir (Zhou ve ark., 2015). CNN, yazılı metin içindeki yerel dilsel örüntülere (örneğin; kelime dizimleri, n-gram benzeri yapılar) odaklanırken LSTM bağlamı, cümleler ve kelimeler arasındaki ardışıklığı yakalamayı hedefler; yani bu örüntülerin bağlam içindeki ilişkilerini öğrenir (Taghipour & Ng, 2016).

Bu modellerin otomatik puanlamadaki işleyişi en genel biçimiyle şöyle açıklanabilir: Öncelikle yazılı performans, derin öğrenme modelinin işleyebileceği sayısal bir temsile dönüştürülür. Bu dönüşüm genellikle her kelimenin belirli bir vektörle temsil edildiği vektör temsili[2] katmanı aracılığıyla gerçekleştirilir (Pennington ve ark., 2014). Böylece metindeki her kelime, anlamsal ve sözdizimsel ilişkilerini koruyan çok boyutlu bir uzayda yer alır. Model bu vektörleri kullanarak kelimeler arasındaki benzerlikleri, örüntüleri ve bağlamsal ilişkileri öğrenebilir.

Bu temsiller daha sonra CNN katmanlarına aktarılır. CNN katmanları, metindeki kısa kelime dizileri veya *n-gram* düzeyindeki örüntüler üzerinde çalışarak yüzeysel fakat güçlü biçimde ayırt edici dilsel özellikleri yakalar. Bu yapı, metni küçük dilsel parçalara ayırarak her bir parçadaki kalıpları tarar ve böylece yazının biçimsel düzeni ile söylemsel organizasyonuna ilişkin ipuçlarını ortaya çıkarır. Örneğin; "in conclusion" (sonuç olarak) veya "as a result of" (-nin sonucu olarak) gibi söylemsel belirteçler, metnin yapısal

[2] Her kelime, modelin dilin anlamını öğrenebilmesi için sayısal bir biçimde vektör temsili (embedding) olarak kodlanır.

olgunluğu ve retorik bütünlüğü hakkında dolaylı bilgi sağlayan örüntülerdir. CNN tabanlı otomatik puanlama modelleri bu tür yüzeysel örüntüleri yakalayabilmektedir (Dong ve ark., 2017). Bu tarama sürecinin sonucunda, metnin yüzey yapısında yer alan *n-gram* düzeyindeki örüntüler otomatik olarak belirlenir ve daha ileri katmanlarda kullanılmak üzere giderek daha soyut ve yüksek düzeyli dilsel özelliklere dönüştürülür (Zhou ve ark., 2016).

Elde edilen yerel özellikler daha sonra LSTM katmanına aktarılır; burada amaç metin düzeyinde daha uzun bağımlılıkların öğrenilmesidir (Tashu ve ark., 2022; Zhou ve ark., 2016). LSTM, RNN'lerin aksine mesafeli bağlamı koruyabilir; yani sözcükler arasındaki uzaklık artsa da aralarındaki bağlam LSTM sayesinde anlaşılabilir (Greff ve ark., 2017). Bu sayede model, cümleler arasında kurulan mantıksal geçişleri, neden-sonuç ilişkilerini ve anlatıdaki bütünlüğü çözümleyebilir (Ghojogh & Ghodsi, 2023). LSTM katmanları, özellikle bireyin metinde bir fikri nasıl geliştirdiğini veya argümanını nasıl yapılandırdığını tanımada etkilidir. CNN katmanlarından gelen yerel örüntü bilgisi ile LSTM'nin bağlamsal hafızası birleştiğinde model hem mikro (kelime-dizi) hem makro (paragraf-bağlam) düzeyde anlamlı dil yapılarını tanıyabilir (Yin ve ark., 2017).

Son aşamada LSTM katmanının çıktısı özetlenir. Bu özetleme işlemi, modelin farklı zaman adımlarından elde ettiği bilgiyi tek bir vektörde topladığı "havuzlama" (pooling) adımıyla gerçekleştirilir. Böylece model, metni genel bir bütün olarak temsil eden tek bir özet vektör oluşturur. Bu vektör daha sonra "yoğun katman"a (dense layer) aktarılır ve burada öğrenilmiş ağırlıklar aracılığıyla metne nihai bir puan verilir. Bu mimari sayesinde model, insan değerlendiricilerin yaptığına benzer şekilde hem yerel dilsel örüntüleri hem de bütünsel anlatı yapısını dikkate alan tutarlı bir puanlama gerçekleştirebilir (Lopez, 2017; Roy & Kumar, 2021). Dolayısıyla derin öğrenme tabanlı yaklaşımlar, önceki kural ve özellik temelli modellerin sınırlı temsil gücünü aşarak metnin biçimsel ve anlamsal boyutlarını bütüncül biçimde değerlendirebilmektedir.

#### *2.3.2.2. Derin Öğrenmeden Üretken Yapay Zekâ'ya Geçiş*

CNN ve LSTM gibi derin öğrenme modelleri, metinleri sayısal temsiller üzerinden işlemeye odaklanırken üretken yapay zekâ yaklaşımları, özellikle *Transformer* mimarisine dayalı modeller, dilin kendine özgü bağlamsal doğasını doğrudan modelleme kapasitesi sunar (Vaswani ve ark., 2017). LSTM ve benzeri sıralı modeller, zaman adımlarındaki bilgi aktarımında sınırlılıklar gösterirken (Liu ve ark., 2021), *Transformer* mimarisi bu sınırlılıkları aşarak paralel işlem gücü ve öbek hâlinde bağlam analizine olanak tanır (Vaswani ve ark., 2017). Dolayısıyla LSTM gibi sıra temelli modellerin öğrenme kapasitesini aşan durumlarda, *Transformer* tabanlı büyük dil modelleri otomatik puanlamada yeni bir paradigma oluşturmuştur.

*Transformer Mimarisi ve Temel İlkeleri*

*Transformer* mimarisi, günümüz ÜYZ modellerinin temel yapıtaşlarından biridir. İlk olarak Encoder–Decoder biçiminde tanımlanmış olan bu mimaride, metin önce bir kodlayıcı (encoder) katmandan geçirilir. Ardından çözümlenmiş temsiller bir çözücüye (decoder) aktarılır ve hedef çıktıya dönüştürülür; yani son aşamada, modelin hedeflediği sonuç (örneğin çeviri, özet veya bu bölüm özelinde puan) üretilir (Vaswani ve ark., 2017). Bu yapı içinde en kritik unsur self-attention mekanizmasıdır: Her bir *token* (kelime ya da sembol), diğer tüm token'lerle dinamik biçimde ilişki kurarak hem yerel (kelimenin metinde kendisine fiziksel olarak yakın olan kelimelerle ilişkisi) hem de bütünsel (kelimenin metnin genelindeki anlam ilişkileri) bağlamları eşzamanlı olarak öğrenebilir (Vaswani ve ark., 2017; IBM, 2024). *Transformer* mimarisinde, sıranın anlamını korumak için positional encoding adı verilen konumsal yerleştirme işlemi uygulanır; bu sayede kelimeler arasındaki bağıntılar sıralı bir şekilde değil, bağlama göre modellenir (OpenReview, 2020). Böylece, geleneksel sıralı modellerin (örneğin LSTM) zamana bağlı bağımlılıkları yakalamadaki sınırlılıklarının ötesine geçilir, bağımlılıklar artık sıraya değil bağlama göre ele alınmış olur (Vaswani ve ark., 2017). Bu durum otomatik puanlama açısından metinlerin yalnızca diziler olarak değil anlam ve bağlam ilişkileriyle bir bütün olarak değerlendirilmesine olanak tanımış ve modellerin dilsel bütünlüğü insan benzeri biçimde çözümleyebilmesini mümkün kılmıştır (Devlin ve ark., 2019).

#### *2.3.2.3. Üretken Yapay Zekâ (ÜYZ) Temelli Otomatik Puanlama Modelleri*

Bahsi geçen özellikleriyle GDM, önceki yaklaşımlara kıyasla çok daha düşük düzeyde özellik mühendisliği ve ön-tanımlı kural gereksinimiyle otomatik puanlama yapmayı mümkün kılmaktadır. Kural temelli yaklaşımlarda, puanlamanın geçerliğini sağlamak için alan uzmanlarınca puanlama kriterlerinin ve kurallarının önceden ayrıntılı biçimde tanımlanması zorunludur. Makine öğrenmesi temelli yaklaşımlarda sistemin öğrenebilmesi için insan puanlayıcıların verdiği puanlardan oluşan etiketli veri setleri kullanılarak bir tahmin algoritması geliştirilir. ÜYZ temelli yaklaşımlarda ise zaman ve emek açısından maliyetli bu süreçlere gerek kalmadan yüksek doğruluk düzeyinde puanlama yapılabilir. GDM'ler, dilin bağlamsal yapısını doğrudan modelleyebildiği için doğru kurgulanmış bir istem (prompt) aracılığıyla, hangi ölçütlere göre değerlendirme yapacağını kavrayıp uygun puanı üretebilmektedir. Böylelikle modelden artık önceden tanımlanmış kuralları veya yüzeysel istatistiksel eşleşmeleri değil geniş bağlam ilişkilerini esas alan bütüncül bir yargıda bulunması beklenmektedir.

Son yıllarda GDM'lerin gelişmesi, otomatik puanlamada geleneksel makine öğrenmesi yaklaşımlarına alternatif olarak istem-tabanlı puanlamayı, yani sıfır-atım (*zero-shot)* veya birkaç-atım (*few-shot)* değerlendirmeyi mümkün kılmıştır. Zero-few shot kavramları aslında otomatik puanlamaya özgü değildir. *Zero-shot* senaryolar modelin yalnızca verilen yönergeye dayanarak istenen bir işlemi yapmasını, *few-shot* senaryolar ise istenen işlemi yapmak için modele birkaç örnek gösterilmesini esas alır (Brown ve

ark., 2020). Xiao ve arkadaşları (2024) ÜYZ temelli otomatik puanlama yöntemlerini rubriksiz *zero-shot,* rubrikli *zero-shot* ve rubrikli *few-shot* olmak üzere üç kısımda ele almıştır (akt; Aydın ve ark., 2025): **(1)** Rubriksiz *zero-shot* yaklaşımında modele yalnızca istem ve puanlanacak yazılı performans sunulur; **(2)** rubrikli *zero-shot* yaklaşımında isteme ek olarak değerlendirmede esas alınacak rubrik de verilir; **(3)** rubrikli *few-shot* yaklaşımında ise rubrikle birlikte birkaç örnek performans ve bu performansların insan puanlayıcılar tarafından verilmiş puanları modele gösterilerek değerlendirme yapılması sağlanır.

*Fine-tuning* (ince ayar) ise istem-tabanlı bu yaklaşımların ötesine geçerek, modelin görev spesifik puanlama ölçütlerini doğrudan öğrenmesini sağlayan ek bir uyarlama süreci sunar (OpenAI, 2023). *Zero-shot* veya *few-shot* senaryolar büyük dil modellerinin genel kapasitesine dayanırken *fine-tuning* belirli bir yazma türüne, rubriğe veya hedef performans özelliklerine ilişkin örüntüleri modelin parametrelerine işler. Bu sayede model puan tahmini yapmakla kalmaz, aynı zamanda puanın gerekçesine ilişkin açıklamalar üretebilecek düzeyde göreve uyarlanmış bir değerlendirme yetkinliği kazanır (Xiao ve ark., 2024). Ancak bu işlem oldukça maliyetlidir.

#### *2.3.2.4. Otomatik Puanlama Performansının Değerlendirilmesi*

Otomatik puanlamaya ilişkin en önemli hususlardan biri, puanlama işleminin performansını-hatasızlığını (accuracy) ortaya koymak; bir diğer ifadeyle geçerliğine ilişkin kanıt elde etmektir. Otomatik puanlamanın performans kıstası ise insan puanlayıcılarla uyum düzeyidir. Bu uyumu ortaya koymada en çok kullanılan istatistik Quadratic Weighted Kappa-*QWK*'dır. *QWK*, iki puanlayıcının (örneğin, insan–insan veya insan–yapay zekâ) aynı dereceli ölçek üzerinde ne kadar tutarlı puan verdiğini ölçen bir uyum katsayısıdır (Williamson ve ark., 2012)**.** Buradaki 'quadratic' kavramı, ağırlıklandırmanın doğrusal değil farkların karesine dayalı olmasını ifade eder. Haliyle puanlayıcılar arasındaki puan farkı büyüdükçe, uyum katsayısındaki düşüş daha belirgin hâle gelir. *QWK'nin* Haberman (2019) tarafından sunulan varyans–kovaryans temelli eşdeğer gösterimi Eşitlik 1'de verilmiştir:

$$QWK = \frac{2\,\mathrm{Cov}(X,Y)}{\mathrm{Var}(X) + \mathrm{Var}(Y) + (E[X] - E[Y])^2} \qquad \textit{Eşitlik 1}$$

Eşitlik 1'de *Cov(X,Y)*, iki puanlayıcının verdiği puanlar arasındaki kovaryansı; *Var(X)* ve *Var(Y)* her bir puanlayıcının puan dağılımlarının varyansını; *E[X]* ve *E[Y]* ise beklenen (ortalama) puanları ifade eder. Payda, iki puanlayıcı arasındaki toplam puanlama değişkenliğini temsil ederken pay bu değişkenlik içinde ortak hareket eden kısmı gösterir. Böylece *QWK* puanlayıcılar arasındaki uyumu, puan farklarının büyüklüğünü de dikkate alarak ±1 arasında değerler alacak biçimde ortaya koyar. Otomatik puanlama performansının yeterli düzeyde olduğunun kabul edilebilmesi için bu değerin 0.70 ve üzerinde olması beklenir (Williamson ve ark., 2012).

Otomatik puanlama uygulamalarında geleneksel Kappa yerine *QWK'nın* tercih edilmesinin nedeni, Kappa'nın nominal veri varsayımı gereği tüm puan farklarını aynı düzeyde uyumsuzluk olarak değerlendirmesidir. Oysa *QWK*, ordinal ölçek yapısını dikkate alarak puan farklarını büyüklüklerine göre ağırlıklandırır. Örneğin bir puanlayıcının 6, diğerinin 7 vermesi görece küçük bir uyumsuzluk iken birinin 2 diğerinin 7 vermesi belirgin bir uyumsuzluk olarak ele alınır. Bu ayrım *QWK* ile ortaya konulabilmektedir.

*QWK*, otomatik puanlama modellerinde insan puanlarıyla uyumu değerlendirmek için en yaygın kullanılan ölçütlerden biridir ve çoğu çalışmada Pearson korelasyonuna benzer biçimde yorumlanır (Cohen, 1968; Williamson ve ark., 2012). Ancak *QWK*'nin değeri yalnızca model başarısından değil puan dağılımlarının ortalama ve varyans eşleşmesinden, ölçeğin değer aralığından ve puanlayıcıların belirli kategorilerde yoğunlaşmasından da etkilenir. Yüksek prevalans (puan kategorilerinden bazılarında aşırı yığılma) veya yanlılık durumlarında *QWK* yapay biçimde yükselebilir ya da düşebilir (Lewis & Casabianca, 2023; Sim & Wright, 2005; Xu & Lorber, 2014). Ayrıca puanlama ölçeği daraldıkça küçük farklılıkların ağırlığı arttığı için *QWK* genellikle daha düşük değerler alır (Brenner & Kliebsch, 1996; Parker ve ark., 2013). İnsan puanlayıcılar arasındaki tutarlılık düşük olduğunda da *QWK*, modelin gerçek performansını olduğundan daha zayıf gösterebilir (Loukina ve ark., 2020). Bu nedenle *QWK* tek başına mutlak bir doğruluk göstergesi olarak alınmamalı; bağlam, dağılımlar ve puanlayıcı güvenilirliği göz önünde bulundurularak yorumlanmalıdır (McCaffrey ve ark., 2025).

### 10.2.4. Karşılaştırma Yargılarıyla Puanlama

Buraya kadar açıklanan puanlama türleri, performansın mutlak bir ölçüte dayanarak puanlanmasını esas alır; her performansın-yanıtın diğerlerinden bağımsız olarak bir değeri vardır ve puanlayıcı (ister insan ister algoritma) bu değeri bir ölçüte dayalı olarak biçebilir. Bununla birlikte mutlak puanlama sürecinde incelenen performans dışında başka performanslar da yapılan puanlamaları etkileyebilir. Bu durumda puanlayıcı, yaptığı puanlamada sadece önceden belirlenmiş ölçütleri değil diğer performansları da bir kıyas olarak kullanmış olur. Laming (2004) bunu "mutlak değerlendirme yoktur; tüm değerlendirmeler birinin diğeriyle karşılaştırılmasına dayanır" sözüyle ifade etmiştir. Analitik puanlamadan yararlanılan kimi durumlarda ise parçaların toplamı, bütünden farklı bir şey ifade edebilmektedir. Yaratıcılığı ölçmeye çalıştığımız bir uygulamada, puanların güvenirliğini artırabilmek için analitik puanlama yaptığımızı düşünelim. Burada analitik puanlama aracının her bir bileşeni özenle hazırlanmış olsa dahi bu parçalardan elde edilen toplam puan her zaman yaratıcılığı yansıtmayabilir (Jones & Davies, 2023).

Thurstone'nin karşılaştırmalı yargı yaklaşımını esas alan karşılaştırmalı yargıyla puanlama (comparative judgement) - (KYP), esas itibariyle model temelli bir yaklaşımdır. Bununla beraber modelin kullandığı verinin üretiliş biçimi ontolojik olarak farklılaştığından KYP'yi bölüm kapsamında ayrı bir başlık şeklinde ele almak daha uygun bulunmuştur.

KYP yönteminde puanlayıcıya iki farklı bireyin performansı sunulur ve puanlayıcıdan yalnızca bütünsel açıdan hangi performansın diğerinden daha üstün olduğuna karar vermesi beklenir. Her performans bu şekilde çok sayıda ikili karşılaştırma sürecinden geçer ve bir matrise kaydedilir. Modelleme ise *Rasch* modeline oldukça yakın bir yapıya sahip olan Bradley-Terry-Luce (BTL) (Bradley & Terry, 1952; Luce, 1959) ile yapılır. KYP'nin çıkış noktası, mutlak bir puana karar vermektense göreli bir yargıda bulunmanın daha kolay olduğu fikridir (Bramley, 2005). Ayrıca iki puanlayıcı belirli bir performansı değerlendirirken aynı puanı vermeyebilir; fakat iki performanstan hangisinin diğerinden "daha iyi" olduğu konusunda uzlaşmaları daha olasıdır (Bramley, 2007; Steedle & Ferrara, 2016). Nitekim ikili karşılaştırmalarla yapılan değerlendirmelerde, puanlayıcıların yalnızca kısa ve yüzeysel bir eğitimle süreci etkili biçimde yürütebildikleri belirtilmektedir (Gürel ve ark., 2025). KYP uygulaması performansların eşleştirilmesi, yetenek kestirimi, güvenirlik ve sonlandırma kuralı şeklinde dört temel bileşen üzerinden açıklanabilir. Bu bileşenlerin her biri aşağıda ayrı bir başlık olarak sunulmuştur.

Yetenek Kestirimi Thurstone (1927) karşılaştırmalı yargı yasasıyla gizil bir ölçek üzerinde nesnelerin göreli konumlarını tahmin etmeye yönelik ilk yaklaşımı ortaya koymuştur. Ardından Bradley ve Terry (1952) ile Luce (1959), karşılaştırmalı yargı verilerinin analizinde lojistik fonksiyonların nasıl kullanılabileceğini göstermiş, Andrich (1978) ise Thurstone'un yaklaşımının Rasch lojistik modeliyle örtüştüğünü ifade etmiştir. Karşılaştırma modeli Eşitlik 2'deki tanımlanır:

$$P_k(a_j) = p(X_j k = 1 | a_j, a_k) = exp(a_j - a_k) / \left(1 + exp(a_j - a_k)\right) \qquad \textit{Eşitlik 2}$$

Burada $P_{\mathrm{k}}(a_{\mathrm{j}})$ ya da $p(X_{\mathrm{jk}} = 1 | a_{\mathrm{j}}, a_{\mathrm{k}})$, $j$ performansının $k$ performansına tercih edilme olasılığını ifade eder. $X_{jk}$=1 durumu, $j$ performansının $k$'ye göre seçildiğini gösterir. Yetenek değerleri $\alpha_j$ ve $\alpha_k$ ilgili performanslara ait logit birimindeki başarı düzeylerini temsil eder. Yetenek kestirimi ise maksimum olabilirlik metoduyla yapılır.

***Performansların Eşleştirilmesi*** konusunda öne çıkan yöntem yarı-rassal eşlemedir. Bu yöntemde her bir performans bir diğeriyle daha önceki eşleştirmeler dışarıda bırakılarak rassal olarak eşleştirilir. İsviçre satrancı olarak anılan yöntemde ise amaç eşleşmeyi uyarlamalı olarak yapabilmektir. Toplam karşılaştırma sayısını azaltırken güvenirliği yüksek tutmayı mümkün kılan başka uyarlamalı yöntemler de bulunmaktadır (ayrıntılar için bkz: Gürel ve ark., 2025).

***KYP'de güvenirlik***, ölçeğin ayırt etme gücünü ifade eden SSR (scale separation reliability) ile hesaplanır. SSR, değerlendiricilerin performans sıralamalarına ilişkin ne düzeyde hemfikir olduğunda ilişkin bilgi sunar. Formülü şu şekildedir: (Gustafsson, 1977):

$$\mathrm{SSR} = (\sigma^{a2} - \mathrm{MSE}) / \sigma^{a2} \qquad \textit{Eşitlik 3}$$

Sonlandırma Kuralı KYP'de çoğunlukla belirli bir karşılaştırma sayısına ya da ortalama karşılaştırmaya dayalı bir sonlandırma kuralı işletilir (Cromptvoets ve ark., 2020). Farklı yetenek düzeylerine sahip performanslara eşit sayıda karşılaştırma yapılması, uç noktalardaki performansların kestirim standart hatasının artmasına yol açmaktadır (Cromptvoets ve ark., 2020; Uysal ve ark., 2024). Bunun önüne geçebilmek içni Verhavert ve arkadaşları (2022), sonlandırma kuralında SSR değerinin kullanılmasını önermiştir. Bu kurala göre belirli bir güvenirlik düzeyine ulaşıldığında süreç durdurulur. Ancak yüksek SSR bile tüm kestirimlerin aynı standart hata düzeyine sahip olmasını garanti etmez. Bu nedenle sonlandırma sürecinin BBT'de olduğu gibi, doğrudan kestirim standart hatasına bağlı olarak yürütülmesi önerilir (Gürel ve ark., 2025).

## 3. SONUÇ

Puanlama yöntemleri, eğitimde ölçme ve değerlendirme sürecinin en kritik bileşenlerinden birini oluşturmaktadır. Ölçme aracının niteliği, ölçülmek istenen özellik, değerlendirmenin amacı ve uygulama koşulları, hangi puanlama yaklaşımının tercih edileceğini belirleyen temel unsurlar olarak görülebilir. Bu nedenle tek bir puanlama yönteminin tüm değerlendirme durumları için en uygun çözüm olduğunu söylemek mümkün değildir. Teknolojik gelişmeler, özellikle otomatik puanlama alanında önemli dönüşümleri beraberinde getirmiştir. Kural ve özellik temelli sistemlerden derin öğrenme ve büyük dil modellerine uzanan bu gelişim, üretimsel becerilerin daha hızlı, tutarlı ve ölçeklenebilir biçimde değerlendirilmesini mümkün kılmaktadır. Bununla birlikte, yapay zekâ temelli sistemlerin sağladığı yüksek performans, puanlama sürecine ve sonuçlarına ilişkin tartışmaları ortadan kaldırmamaktadır. Bu nedenle otomatik puanlama sistemlerinin, özellikle geçerlik konusunda vermesi gereken önemli sınavlar vardır. Diğer yandan, karşılaştırma yargılarıyla puanlama gibi alternatif yaklaşımlar, özellikle bütüncül değerlendirmenin ön planda olduğu performans görevlerinde önemli avantajlar sunmaktadır. Yapay zekâyla karşılaştırmalı yargılamanın birlikte kullanılmasına yönelik çalışmalar, gelecekte hibrit puanlama sistemlerinin yaygınlaşabileceğine işaret etmektedir. Böyle sistemlerin, psikometrik modellerle yapay zekânın hız ve ölçeklenebilirliğini birleştirerek değerlendirmelerin güvenirliğine ve geçerliğine katkı sunabileceği düşünülmektedir. Sonuç olarak, puanlama yöntemlerindeki gelişim yalnızca teknolojik ilerlemelerle değil, aynı zamanda ölçme kuramındaki gelişmelerle de şekillenmektedir. Gelecekteki çalışmaların, farklı puanlama yaklaşımlarını karşılaştırmalı olarak incelemesi ve bu yöntemlerin geçerlik, güvenirlik, adillik ve uygulanabilirlik boyutlarını birlikte ele alması, eğitimde daha nitelikli değerlendirme sistemlerinin geliştirilmesine önemli katkılar sağlayacaktır.

## KAYNAKÇA | REFERENCES

Andrich, D. (1978). Relationships between the Thurstone and Rasch approaches to item scaling. *Applied Psychological Measurement*, *2*(3), 451–462. https://doi.org/10.1177/014662167800200319

Benton, T. (2018, Ekim). *Exploring the relationship between optimal methods of item scoring and selection and predictive validity*. Paper presented at the 19th Annual Conference of the Association for Educational Assessment in Europe, Arnhem-Nijmegen, The Netherlands.

Benton, T. (2021). Comparative judgement for linking two existing scales. *Frontiers in Education*, *6*, 775203. https://doi.org/10.3389/feduc.2021.775203

Bradley, R. A. & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, *39*(3–4), 324–345. https://doi.org/10.1093/biomet/39.3-4.324

Bramley, T. (2005). A rank-ordering method for equating tests by expert judgment. *Journal of Applied Measurement*, *6*(2), 202–223.

Bramley, T. (2007). Paired comparison methods. P. Newton, J. Baird, H. Goldstein, H. Patrick & P. Tymms (Ed.), *Techniques for monitoring the comparability of examination standards* (s. 246–294) içinde. Qualifications and Curriculum Authority.

Brenner, H. & Kliebsch, U. (1996). Dependence of weighted kappa coefficients on the number of categories. *Epidemiology*, *7*(2), 199–202. https://doi.org/10.1097/00001648-199603000-00016

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Amodei, D. & Ryder, N. (2020). *Language models are few-shot learners*. In Advances in Neural Information Processing Systems.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Educational and Psychological Measurement*, *33*(1), 37–46. https://doi.org/10.1177/001316446802800104

Crompvoets, E. A. V., Béguin, A. A. & Sijtsma, K. (2020). Adaptive pairwise comparison for educational measurement. *Journal of Educational and Behavioral Statistics*, *45*(3), 316–338. https://doi.org/10.3102/1076998619890589

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (s. 4171–4186) içinde. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Dikli, S. (2006). An overview of automated scoring of essays. *Journal of Technology, Learning, and Assessment, 4*(3). https://files.eric.ed.gov/fulltext/EJ843855.pdf

Dong, F., Zhang, Y. & Yang, J. (2017). Attention-based recurrent convolutional neural network for automatic essay scoring. Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017) (s. 153–162) içinde. https://frcchang.github.io/pub/conll17.fei.pdf

Ghojogh, B. & Ghodsi, A. (2023). Recurrent neural networks and long short-term memory networks: Tutorial and survey. arXiv. https://arxiv.org/abs/2304.11461

Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R. & Schmidhuber, J. (2017). LSTM: A search space odyssey. *IEEE Transactions on Neural Networks and Learning Systems, 28*(10), 2222–2232. https://doi.org/10.1109/TNNLS.2016.2582924

Gustafsson, J.-E. (1977). The Rasch model for dichotomous items: Theory, applications and a computer program. Göteborg University.

Gürel, S., Şahin, M. D., Uysal, İ., İbileme, A. İ. & Gündüz, T. (2025). Karşılaştırma yargılarıyla puanlamada uyarlamalı eşleme ve standart hata sonlandırma kuralı: Yazma becerisinin ölçülmesine yönelik bir uygulama. *Eğitim ve Bilim, Ek Sayı 1,* 93–110.

Haberman, S. J. (2019). *Measures of agreement versus measures of prediction accuracy* (Research Report No. RR-19-20). ETS. https://doi.org/10.1002/ets2.12258

IBM. (2024, Ocak 10). What is self-attention? IBM Think. https://www.ibm.com/think/topics/self-attention

Jones, I. & Davies, B. (2023). Comparative judgement in education research. *International Journal of Research & Method in Education, 47*(2), 170–181. https://doi.org/10.1080/1743727X.2023.2242273

Jurafsky, D. & Martin, J. H. (2023). Speech and language processing (3. baskı). Draft online edition.

Kula Kartal, S. (2021). Performans görevlerine ve portfolyoya dayalı durum belirleme. M. D. Şahin (Ed.), *Eğitimde ölçme ve değerlendirme* (s. 175-206) içinde. Nobel Akademi.

Kurz, T. B. (1999). *A review of scoring algorithms for multiple-choice tests* (Report No. ED 428 076). EDRS Publications.

Laming, D. (2003). *Human judgment: The eye of the beholder*. Thomson Learning.

Lewis, C. & Casabianca, J. M. (2023). On the misuse of quadratic weighted kappa in automated scoring research. *Educational Measurement: Issues and Practice*, *42*(1), 60–72. https://doi.org/10.1111/emip.12552

Liu, H., Shen, Y. & Chen, <J. (2021). Understanding the limitations of LSTMs for long sequences. *IEEE Access, 9*, 63785–63797. https://doi.org/10.1109/ACCESS.2021.3074639

Lopez, M. M. (2017). *Deep learning applied to NLP*. arXiv preprint. https://doi.org/10.48550/arXiv.1703.03091

Loukina, A., Madnani, N., Zechner, K. & Cahill, A. (2020). *Using simulated data to examine metrics for automated scoring evaluation*. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (s. 6072–6082) içinde. https://doi.org/10.18653/v1/2020.acl-main.540

Luce, R. D. (1959). *Individual choice behaviour: A theoretical analysis*. John Wiley & Sons.

Manning, C. D., Raghavan, P. & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University.

McCaffrey, D. F., Casabianca, J. M. & Johnson, M. S. (2025). Measuring the accuracy of true score predictions for AI scoring evaluation. *Journal of Educational Measurement*. Advance online publication. https://doi.org/10.1111/jedm.7001

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013). Efficient estimation of word representations in vector space. In NIPS Workshop.

OpenAI. (2023). Fine-tuning GPT models. https://platform.openai.com/docs/guides/fine-tuning

OpenReview. (2020). On position embeddings in BERT. https://openreview.net/pdf/be0283e323f1b118c975dbc46f7f75c59b467fe0.pdf

Özdemir, D. (2003). Çoktan seçmeli testleri puanlama yöntemlerine bir bakış [An overview of methods for scoring multiple choice tests]. *Eğitim Araştırmaları Dergisi, 4*(12), 121–122.

Page, E. B. (1966). The imminence of... grading essays by computer. *Phi Delta Kappan, 47*(5), 238–243. http://www.jstor.org/stable/20371545.

Parker, P., May, L. & Stahl, J. (2013). The impact of score scale length on reliability and agreement indices. *Journal of Applied Measurement, 14*(2), 123–134.

Pennington, J., Socher, R. & Manning, C. D. (2014). *GloVe: Global vectors for word representation*. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (s. 1532–1543) içinde. Association for Computational Linguistics.

Rowley, G. L. & Traub, R. E. (1977). Formula scoring, number-right scoring, and test-taking strategy. *Journal of Educational Measurement*, *14*(1), 15–22.

Roy, P. K. & Kumar, A. (2021). Convolutional neural network for text: A stepwise working guidance. Proceedings of Yukthi 2021 – The International Conference on Emerging Trends in Engineering içinde. https://www.researchgate.net/publication/356795494

Shermis, M. D. & Burstein, J. (2013). Handbook of automated essay evaluation. Routledge.

Sim, J. & Wright, C. C. (2005). The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy, 85*(3), 257–268. https://doi.org/10.1093/ptj/85.3.257

Stanley, J. C. & Wang, M. D. (1970). Weighting test items and test-item options: An overview of the analytical and empirical literature. *Educational and Psychological Measurement, 30*(1), 21–35.

Taghipour, K. & Ng, H. T. (2016). A neural approach to automated essay scoring. Proceedings of the Conference on Empirical Methods in Natural Language Processing (s. 1882–1891) içinde.

Tashu, T. M., Maurya, C. K. & Horváth, T. (2022). *Deep learning architecture for automatic essay scoring*. arXiv preprint. https://arxiv.org/abs/2206.08232

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34*(4), 273–286. https://doi.org/10.1037/h0070288

Uysal, İ., Gürel, S., Şahin, M. D., İbileme, A. İ. & Yıldırım, Y. (2024, Eylül). *Açık uçlu maddelerin karşılaştırmalı yargıyla puanlanmasında sabit sayıda uyarlamalı ve rassal eşlemeye dayalı bir simülasyon çalışması*. 9. Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongresi, Anadolu Üniversitesi, Eskişehir.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. (2017). Attention is all you need. I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan & R. Garnett (Ed.), Advances in Neural Information Processing Systems (Vol. 30). Curran Associates içinde. https://doi.org/10.5555/3295222.3295349

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. (2017). *Attention is all you need.* arXiv preprint arXiv:1706.03762. https://doi.org/10.48550/arXiv.1706.03762

Verhavert, S., Furlong, A. & Bouwer, R. (2022). The accuracy and efficiency of a reference-based adaptive selection algorithm for comparative judgment. *Frontiers in Education, 6*, 785919. https://doi.org/10.3389/feduc.2021.785919

Williamson, D. M., Xi, X. & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice, 31*(1), 2–13. https://doi.org/10.1111/j.1745-3992.2011.00223.x

Xu, R. & Lorber, M. F. (2014). Interrater agreement statistics with skewed data: Evaluation of weighted kappa. *Psychological Methods*, *19*(4), 524–539. https://doi.org/10.1037/met0000014

Yaşar, M., Kula Kartal, S. & Aybek, E. C. (2021). Scoring methods for multiple choice tests: How does the item difficulty weighted scoring change students' test results? *Bartın University Journal of Faculty of Education*, *10*(2), 309–324. https://doi.org/10.1016/buefad.878504

Yin, W., Kann, K., Yu, M. & Schütze, H. (2017). *Comparative study of CNN and RNN for natural language processing*. arXiv preprint arXiv:1702.01923. https://arxiv.org/abs/1702.01923

Yurdugül, H. (2010). Farklı madde puanlama yöntemlerinin ve farklı test puanlama yöntemlerinin karşılaştırılması. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *1*(1), 1–8.

Zhou, C., Sun, C., Liu, Z. & Lau, F. C. M. (2015). *A C-LSTM neural network for text classification*. arXiv. https://arxiv.org/abs/1511.08630

Zhou, C., Sun, C., Liu, Z. & Lau, F. C. M. (2016). *Text classification improved by integrating bidirectional LSTM with two-dimensional max pooling*. https://aclanthology.org/C16-1329.pdf

## YAZAR HAKKINDA | ABOUT AUTHOR

**DOÇ. DR. MURAT DOĞAN ŞAHİN**

**Anadolu Üniversitesi | mdsahin[at]anadolu.edu.tr**

**ORCID: 0000-0002-2174-8443**

Doç. Dr. Murat Doğan ŞAHİN, 1984 yılında Kütahya'da doğmuş, ilk ve ortaöğrenimini burada tamamlamıştır. 2006 yılında Hacettepe Üniversitesi Fen ve Matematik Alanları Eğitimi Bölümünden mezun olmuş; 2011 yılında aynı üniversitede Ölçme ve Değerlendirme alanında yüksek lisansını, 2017 yılında ise doktorasını tamamlamıştır. 2012–2017 yılları arasında Anadolu Üniversitesi Açıköğretim Fakültesi Test Araştırma Biriminde uzman olarak görev yapmıştır. 2018 yılından bu yana Anadolu Üniversitesi Eğitim Fakültesi Eğitimde Ölçme ve Değerlendirme Anabilim Dalında öğretim üyesi olarak çalışmalarını sürdürmektedir. Araştırmaları ağırlıklı olarak açık uçlu yanıtların puanlanmasına yönelik yenilikçi yaklaşımlar, yapay zekâ destekli değerlendirme sistemleri, bireye uyarlanmış testler ve dijital ölçme uygulamaları üzerine odaklanmaktadır.

**ASSOC. PROF. DR. MURAT DOĞAN ŞAHİN**

**Anadolu University | mdsahin[at]anadolu.edu.tr**

**ORCID: 0000-0002-2174-8443**

Associate Professor Dr. Murat Doğan ŞAHİN was born in Kütahya in 1984 and completed his elementary and secondary education there. He graduated from the Department of Science and Mathematics Education at Hacettepe University in 2006; he completed his master's degree in Measurement and Evaluation at the same university in 2011 and his Ph.D. in 2017. Between 2012 and 2017, he served as a specialist at the Test Research Unit of the Faculty of Open Education at Anadolu University. Since 2018, he has been continuing her work as a faculty member in the Department of Measurement and Evaluation in Education at the Faculty of Education, Anadolu University. His research focuses primarily on innovative approaches to scoring open-ended responses, artificial intelligence-supported assessment systems, individualized tests, and digital assessment applications.