T.C.

EGE ÜNİVERSİTESİ

SOSYAL BİLİMLER ENSTİTÜSÜ

Eğitim Bilimleri Anabilim Dalı

Eğitimde Ölçme ve Değerlendirme Bilim Dalı

TEST GELİŞTİRME SÜRECİNDE MADDE SEÇİMİ İÇİN HİYERARŞİK PUANLAYICI MODEL: SİNYAL ALGILAMA PUANLAYICI MODELİNİN UYARLAMASI

YÜKSEK LİSANS TEZİ

Elçin MANAP

JÜRİ ÜYELERİ

Doç. Dr. T. Oğuz BAŞOKÇU (Danışman)

Prof. Dr. Nuri DOĞAN

Doç. Dr. Tuncay ÖĞRETMEN

İZMİR-2017

BİLDİRİM

EGE ÜNİVERSİTESİ SOSYAL BİLİMLER ENSTİTÜSÜ

ETİK KURALLARINA UYGUNLUK BEYANI

Ege Üniversitesi Sosyal Bilimler Enstitüsü Müdürlüğüne sunduğum "**Test Geliştirme Sürecinde Madde Seçimi İçin Hiyerarşik Puanlayıcı Model: Sinyal Algılama Puanlayıcı Modelinin Uyarlanması**" adlı **yüksek lisans tezinin** tarafımdan bilimsel, ahlak ve normlara uygun bir şekilde hazırlandığını, tezimde yararlandığım kaynakları bibliyografyada ve dipnotlarda gösterdiğimi onurumla doğrularım.

21.08.2017

Elçin Manap

T.C.EGE ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ

YÜKSEK LİSANS

TEZ SAVUNMA TUTANAĞI

ÖĞRENCİNİN

Adı Soyadı : **Elçin MANAP**

Numarası : **92150000232**

Anabilim Dalı : **Eğitim Bilimleri**

Tez Başlığı (Türkçe) : **Test Geliştirme Sürecinde Madde Seçimi İçin Hiyerarşik Puanlayıcı Model: Sinyal Algılama Puanlayıcı Modelinin Uyarlanması**

Tez Başlığı (İngilizce) : **The Hierarchical Rater Model in Item Selection of Test Development Process: Adaption of Signal Detection Rater Model**

Tez Savunma Tarihi : **21.08.2017**

Tez Başlığı Değişikliği Varsa Yeni Başlık:

JÜRİ ÜYELERİ

Jüri Başkanı

Unvan, Adı, Soyadı : Doç. Dr. T. Oğuz BAŞOKÇU

Karar : ● Başarılı O Başarısız O Düzeltme

İmza :

Jüri Üyesi

Unvan, Adı, Soyadı : Prof. Dr. Nuri DOĞAN

Karar : ● Başarılı O Başarısız O Düzeltme

İmza :

Jüri Üyesi

Unvan, Adı, Soyadı : Doç. Dr. Tuncay ÖĞRETMEN

Karar : ● Başarılı O Başarısız O Düzeltme

İmza :

TEZ HAKKINDA JÜRİNİN GENEL GÖRÜŞÜ

(Jüri Başkanı Tarafından Doldurulacaktır)

Tez savunması sonucunda öğrenci tarafından hazırlanan çalışma;

Oybirliğiyle ●

Oy çokluğuyla O

Başarılıdır	Düzeltilmelidir	Başarısızdır
X		

İÇİNDEKİLER

TABLOLAR

ŞEKİLLER

KISALTMALAR

Orijinal	Kısaltma	Çeviri	Kısaltma
Biased Downward		Negatif Yanlılık	
Accumulated Fisher Information	AFI	Fisher Toplamlı Bilgi Kriteri	FTBK
Bayesian Approach to Estimation		Tahminde Bayesian yaklaşımları	
Central Tendency		Merkezi Eğilim	
Constructed Response	CR	Yapılandırılmış Yanıt	YY
Expected A Posteriori Estimation	EAP	Beklenen Sonsal Kestirim	BSK
Essay		Açık Uçlu Madde Yanıtı	
Generalized Partial Credit Model	GPCM	Genelleştirilmiş Kısmi Puanlama Modeli	GKPM
Graded Response Model		Dereceli Yanıt Modeli	
Graduate Record Examination	GRE	Lisansüstü Kayıt Sınavı	LKS
Hierarchical Rater Model	HRM	Hiyerarşik Puanlayıcı Modeli	HPM
Hierarchical Rater Model-Signal Detection Theory	HRM-SDT	Hiyerarşik Puanlayıcı Modeli - Sinyal Algılama Teorisi	HPM-SAK
Higher-order Factor Model		Daha Yüksek Dereceli Faktör Modeli	
Item Response Theory	IRT	Madde Tepki Kuramı	MTK
Item Step Parameters		Madde Adım Parametresi	
Markov Chain Monte Carlo Methods	MCMC	Markov Chain Monte Carlo Metodu	MCMC
Maximum Likelihood Estimation	MLE	Maksimum Olasılık Tahmini	MOT
National Assessment of Educational Progress	NAEP	Ulusal Eğitsel Gelişme Değerlendirmesi	UEGD
Posterior Mode Estimation	PME	Posterior Mod Tahminleri	PMT
Programme for International Student Assessment	PISA	Uluslararası Öğrenci Değerlendirmesi Programı	UÖDP
Rater		Hakem / Puanlayıcı	
Rater Precision		Hakem Hassasiyeti	
Rater Heverity Parameter		Hakem Katılık Parametresi	
Raters' Strict or Lenient		Hakemin Serlik ya da Cömertliği	
Restriction of the Range		Ranj Daralması	
Scholastic Aptitude Test	SAT	Okul Yetenek Testi	OYT
Signal Detection Model	SDM	Sinyal Algılama Modeli	SAM
Signal Detection Theory	SDT	Sinyal Algılama Kuramı	SAK
Trends in International Mathematics and Science Study	TIMSS	Uluslararası Matematik ve Bilim Araştırmalarında Eğilimler	UMBAE

1. GİRİŞ

Tez çalışmasının bu bölümde problem durumu hakkında bilgi verilmiştir. Madde Tepki Kuramı (MTK), Sinyal Algılama Teorisi (SAK) Modeli, Hiyerarşik Puanlayıcı Modeli (HPM) ve Hiyerarşik Puanlayıcı Modeli - Sinyal Algılama Teorisi (HPM-SAK) Modeli ile ilgili genel bilgiler sunulmuştur. Test geliştirme süreci ve bu süreçteki madde değerlendirme yöntemlerinden bahsedilmiştir. Amaç ve önem, sınırlılıklar ve ilgili araştırmalar üzerinde durulmuştur.

1.1. Problem Durumu

İnsana ait yetenek, başarı, tutum gibi örtük özellikleri ölçmek psikometrinin temel problemidir. Özellikle 20.yüzyıldaki gelişmelerden sonra örtük özellikleri ölçmek için çeşitli yöntemler geliştirilmiş ve gözlenemeyen özellikleri ölçmede önemli gelişmeler kaydedilmiştir. Bu gelişmelerin en önemli sonucu örtük bir özelliği ölçmeye yönelik ölçme aracı geliştirme sürecinin çeşitli standartlara ve teknik prosedürlere dayandırılmasıdır. Bu konuda Downing ve Haladyna (2006), Gronlund (1982) ve Irvine & Kyllonen (2013) gibi pek çok psikometrist çeşitli çalışmalar yapmıştır.

Eğitimde de başarının ölçülmesi özellikle öğrenci yeteneği, eğitim programı, eğitim çıktıları ve ihtiyaçları gibi birçok alan için oldukça önemlidir. Bu noktada ölçme aracı geliştirme süreci ölçülmek istenen özelliğe ait ölçümlerin yeterli güvenirlikte ve geçerlikte olmasının önemli unsurlarından biridir. Test geliştirme sürecine yönelik bütün adımların içinde, ölçme aracının temel yapı taşını oluşturan test maddesi, en önemli basamağı oluşturmaktadır.

Bir maddenin kalitesini belirlemede uygulama sonrası için geliştirilen pek çok istatistiksel yöntem bulunmakla birlikte bir test maddesi pilot uygulama öncesinde ancak uzman kararlarına dayalı olarak değerlendirilebilmektedir. Bu durum özellikle geniş ölçekli sınavlar için oldukça kritik bir değer taşımaktadır. Bu süreçlerde genellikle soru yazarı

olarak alan uzmanları kullanmak ve soru değerlendirme konusunda yeterliğe sahip hakem grupları ile aday maddelerin değerlendirilip pilot uygulamalar için seçim yapmak gibi yöntemler tercih edilmektedir (ETS, PISA, TIMSS).

Hakem kararlarına dayalı ölçümlerde bu kararların güvenirliğini belirlemeye yönelik çeşitli istatistiksel yöntemler kullanılmaktadır. Bu yöntemlerin geneli açık uçlu ya da yapılandırılmış yanıtlı (YY) maddelerin puanlamasında hakem yargılarına ihtiyaç duyulduğu durumlar için geliştirilmiştir.

Bu araştırmada HPM-STD model öğrencilerin öğrenci performansını puanlamada hakem etkisini belirlemek yerine, pilot uygulaması yapılamayan (ÖSYM vb tarafından uygulanan) sınavlar için test geliştirme sürecinde uzaman kararlarıyla madde seçiminin farklı bir istatistiksel yöntemle incelenmesi amacıyla kullanılmaktadır.

HPM-SAK model, hakem puanlaması gereken maddelerin, puanlama güvenirliğini ve geçerliğini arttırmak için hem hakem yanlılığını hem de hakem kalitesini birer faktör olarak kullanan bir puanlama modelidir. Bu model özellikle açık uçlu soruların puanlanması konusunda çeşitli araştırmalarda kullanılmıştır (Decarlo, 2008). YY maddeleri bulunduran testlerde öğrencilerin cevaplarının içeriğindeki değişkenlik, puanlama esnasında puanlayıcının standartlardan ayrılmasına sebep olabilmektedir. Bu yüzden, bu testlerin puanlamasının zor ve zaman alıcı olmasının yanı sıra, tutarsız olma eğilimi de vardır. Coffman (1971), yaptığı araştırmalar sonucunda, farklı puanlayıcıların aynı kâğıdı birbirinden bağımsız şekilde değerlendirdiğinde farklı puanlar verdiğini; aynı zamanda bir puanlayıcının belli bir zaman geçtikten sonra aynı kâğıda önceki verdiği puandan farklı bir puan verdiğini ortaya koymuştur. Puanlamadaki bu tutarsızlıkların sebebini kesin puanlama kurallarının olmayışının ortaya çıkardığı hakem yanlılıkları ve hakemlerin sertlik-cömertlik seviyelerindeki farklılıklar ile açıklamıştır. Açık uçlu, kısa cevaplı ve yapılandırılmış yanıtlı sorularda hakemlerin sertliği ve yanlılığı ile birlikte hakem puanlarının güvenirlik düzeyleri bu ölçümlerin geçerliğini etkilemektedir. Bu konuda Genelleştirilmiş Kısmi Puanlama Modeli (Muraki, 1992) veya Rasch Modeli (Linacre, 1989) temelli modellerin hakem güvenirliğine ilişkin istatistiksel yaklaşımları bulunmaktadır. Bununla birlikte Yao Hakem Etki Modeli gibi hakem katılığını

hesaplayan yaklaşımlar da geliştirilmiştir. HPM-SAK model ise hakem etkileri, hakem katılığı ve hakem niteliği parametrelerini hesaplayarak modeli daha bütüncül bir perspektifle analiz etmeyi amaçlamaktadır.

Araştırmada; kısa cevaplı, açık uçlu ve boşluk doldurma gibi hakem puanlamalarına dayanan soruların hakem yanlılığından arındırılarak puanlanmasına yönelik geliştirilmiş olan HPM-SAK model, madde seçiminde madde kalitesinin belirlenmesi amacıyla kullanılmıştır.

1.2. Hiyerarşik Puanlayıcı Modeli- Sinyal Algılama Teorisi (HPM-SAK)

Açık uçlu sorular, yapılandırılmış yanıtlı ve kısa cevaplı maddeler Ulusal Eğitsel Gelişme Değerlendirmesi (UEGD), TIMSS ve PISA gibi çeşitli ulusal ve uluslararası değerlendirmelerde kullanılmaktadır. Çoktan seçmeli maddelerden farklı olarak bu maddelere verilen yanıtlar doğru veya yanlış olarak puanlanmak zorunda değildir (DeCarlo, 2011). Bu formattaki maddelerin puanlanması için hakemlere ihtiyaç duyulduğundan, bu ölçümlere fazladan bir hata faktörü daha karışmaktadır. Hakem puanları ile oluşan bu etkilerle baş etmek amacıyla farklı yaklaşımlara dayanan çeşitli yöntemler geliştirilmiştir. Bu yaklaşımlar ile ilgili temel problem, "sınava giren kişinin yeterliliğinin daha kesin ölçümü, sınava giren kişiye daha fazla madde vermek yerine daha fazla hakem kullanmasıyla basitçe elde edilebileceğini" öne sürülmesidir (Lawrence T. DeCarlo et al., 2011). Bu iddia, Mariano (2002) tarafından Fisher Toplamlı Bilgi Kriteri (accumulated Fisher information) açısından incelenmiş ve Patz (1996) ve Patz, Junker, Johnson ve Mariano (2002) tarafından da şu şekilde ifade edilmiştir: "Madde başına düşen hakem sayısı arttıkça, MTK Facets modelleri, sınava giren kişinin örtük yeterliğinin (θ) sınırsız kesinlikte ölçümünü verebilecek gibi görünür".

Yukarıdaki problem, bir hakemin, sınava giren kişinin açık uçlu maddeye verdiği yanıtının kalitesi hakkında bilgi verdiğini değil, daha çok, bir sınava giren kişinin yeterliliği hakkında doğrudan bilgi verdiğini kabul etmektedir.

Patz (1996) tarafından geliştirilen Hiyerarşik Puanlayıcı Modeli (HPM) yukarıda belirtilen sorunun önüne geçmeyi amaçlamaktadır. Patz'a göre daha fazla hakemin sağlanması, belirli bir yanıtın hangi kategoriye ait olduğuna dair daha fazla bilgi sağlarken, doğrudan sınava giren kişinin yeterliği hakkında bilgi vermemektedir, bu sebeple sadece daha fazla hakem kullanılarak, yeterlik tahminlerinin kesin olarak elde edilmesi mümkün görülmemektedir.

Patz (1996) ve Patz, Junker, Johnson & Mariano (2002) tarafından ortaya atılan HPM, verilerin hiyerarşik yapısını açıkça ortaya koymaktadır. HPM, modelin ilk düzeyi (hakem modeli) için bir sinyal algılama modeli (SAM) ve ikinci düzey için bir MTK modeli (madde modeli) kullanmaktadır. Bu model ile verilerde hiyerarşik bir yapı söz konusu olur: İlk aşamada, hakemlerin puanı bir açık uçlu maddeye verilen yanıtın ait olduğu "gerçek" kategorinin sıralı göstergeleridir, ikinci aşamada ise, örtük kategoriler, sınava giren kişinin yeterliliğinin sıralı göstergeleridir.

Aslında, Mariano (2002), Patz ve diğerleri (2002), HPM için, yeterlik tahminlerinin standart hatalarının asla, 2. Düzey MTK modelinde gerçek kategorilerin kullanılmasıyla elde edilenlerden daha küçük olamayacağını göstermiştir (Sınırsız sayıda hakem için veya mükemmel algılama için). Bununla birlikte, HPM'nin Düzey 1'de kullanılan modeli için bazı sınırlamalar bulunmaktadır. Örneğin, Patz ve diğerleri (2002), hakem ayırt ediciliğinin yüksek olması durumunda, hakem katılık parametresinin tahminlerini elde etmekte sorunlar ortaya çıktığını belirtmiştir (katılık parametresi, bir hakemin sertlik ya da cömertlik düzeyini göstermektedir). Bir diğer kısıtlama ise modelin, gerçek uygulama verilerinde sıklıkla rastlanan diğer hakem etkileri de olmasına karşın, yalnızca cömertlik ve sertlik anlamındaki hakem etkilerine izin vermesidir. Bununla birlikte, geleneksel sinyal algılama teorisinin (SAK), örtük sınıfların saptanmasıyla ilgili durumlara genişletilmesine dayanan bir model kullanılırsa, bu problemlerin ortaya çıkmayacağı söylenmektedir (Decarlo, 2002, 2005, 2008).

Bu durum yaklaşımın, HPM-SAK modeli olarak tanımlanan HPM'nin ilk düzeyi için kullanıldığında avantajlar sunduğunu göstermektedir. Örneğin, örtük sınıf SAK modeli, basit sertlik veya cömertliğinin ötesinde gerçek uygulama verilerinde görünen çeşitli

hakem etkileri ile başa çıkabilmektedir. Örtük sınıf SAK modelin, örtük kategorik bir tahmin ile genelleştirilmiş doğrusal bir model olduğu göz önüne alındığında, standart yazılımlarda uygulanması da kolaydır. Bu yaklaşım aynı zamanda psikoloji alanında yaygın olarak kullanılan SAK'ın iyi kurulmuş bir sistemini, HPM hakem modeline taşımaktadır. Bu modelin iki temel düzeyi SAK ve MTK kullanılarak oluşturulmuştur.

1.2.1. Madde Tepki Kuramı (MTK)

HPM-SAK'ın ikinci düzeyi Madde Tepki Kuramına (MTK) dayanmaktadır. Madde Tepki Kuramı bir kişinin testte gösterebileceği performansın bir ya da daha fazla özellik sayesinde kestirilebileceğini varsaymaktadır. Bu kestirimi yaparken de bireylerin gözlenen performansları ile performansa ait yetenekler ve özellikler arasındaki ilişkiden yararlanmaktadır (Baker, 2001). Madde puanının, θ vektörü üzerindeki regresyonu "madde karakteristik fonksiyonu" olarak adlandırılmaktadır. Modelde bireyin maddeye ilişkin performansı ile ölçülen özelliğin matematiksel ilişkisi "madde karakteristik eğrisi" ile verilmektedir. Ölçülen örtük özellik tek boyutlu olduğunda fonksiyonu oluşturan parametreler sabit hale gelir ve madde karakteristik eğrisini oluşturmaktadır. Bu eğri belli yetenek düzeyindeki kişinin bir maddeyi doğru yanıtlama olasılığını vermektedir.

1.2.1.1. MTK'ya Göre Madde Parametreleri

MTK'ya göre madde parametreler aşağıdaki gibi özetlenebilir (Baker, 2001);

- *a parametresi (Madde ayırt ediciliği):* Madde karakteristik eğrisinin eğimidir. Maddenin ne kadar ayırt edici olduğuna dair bilgi sağlar, b noktasındaki eğimdir.
- *b parametresi (Madde güçlüğü):* Bir maddenin %50 olasılıkla doğru yanıtlanması için gereken yetenek düzeyidir. Bu durumda b değeri arttıkça maddeyi yanıtlamak için gerekli yetenek düzeyi artar yani madde zorlaşır.
- *c parametresi:* Madde karakteristik eğrisinin y-eksenini kestiği noktadır. Düşük yetenek düzeyindeki bireylerin maddeyi doğru yanıtlama olasılığını verir. Şans başarısı olarak da yorumlanabilir.

Bu parametrelere ek olarak "madde bilgi fonksiyonu" maddenin ölçülen özelliğe dair ne kadar bilgi taşıdığına ilişkin bir tanımlamadır. "Test bilgi fonksiyonu" ise testteki tüm

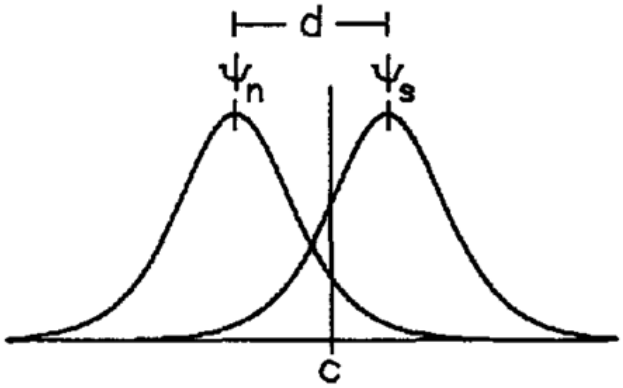
maddelerin madde bilgi fonksiyonlarının toplamıdır. Bu fonksiyon da bir testin ölçülen özelliğe ilişkin ne kadar bilgi sağladığını tanımlamaktadır.

1.2.2. SAK Model

Sinyal algılama teorisi (SAK), istatistiksel karar teorisinin mühendislik problemlerine, özellikle de gürültüye gömülü bir sinyalin algılanmasına, uygulanması olarak ortaya çıkmıştır. Teorinin algılama, tanıma ve ayırt etme konusundaki psikofiziksel çalışmalarla ilişkisi, Tanner and Swets (1954) ve diğerleri tarafından ortaya atılmıştır (Green & Swets, 1966). SAK son yıllarda psikoloji alanındaki çok çeşitli araştırmalara (Gescheider, 1997; N A Macmillan & Creelman, 1991; R J Patz & Junker, 1997; McNicol, 2005; Swets, 1986) ve diğer alanlara, örneğin tıp araştırmalarına, hava durumu tahminine, anket araştırmasına (Green & Swets, 1988; Swets, 2014) ve pazarlama araştırmasına (Singh & Churchill, 1986) uygulanmıştır.

Sinyal algılama teorisinin uygulamaları (Green & Swets, 1966), genellikle tıbbi görüntüleme çalışmasındaki slaytlar veya hafıza araştırmasında kullanılan kelimeler gibi denemeler üzerine farklı maddelerin sunumunu içermektedir. Bir sinyal veya gürültü olmaktan öte, maddelerin kendileri için özel olan faktörlerinin gözlemcilerin tepkilerini etkileyebileceği uzun süredir bilinmektedir (Clark, 1973) ve bu nedenle 'madde etkileri' mevcut olabilmektedir (Freeman, Heathcote, Chalmers, & Hockley, 2010; Morey, Pratte, & Rouder, 2008; Pratte, Rouder, & Morey, 2010; Rouder & Lu, 2005; Rouder et al., 2007).

Bu yaklaşımda, bir madde etkisini temsil etmek için bir örtük sürekli değişken (yani bir madde faktörü) kullanılmaktadır. Bununla birlikte, örtük değişkenin SAK modeline nasıl girdiği tam olarak, durumun nasıl tanımlandığına bağlıdır. Örneğin, madde efekti madde 'zorluğu' olarak düşünülürse, ayırt edicilik parametresini etkileyen örtük bir 'zorluk' değişkeni ortaya çıkabilir; örtük değişken, ayırt edicilik parametresini etkiler çünkü daha zor maddeleri ayırt etmenin daha güç olduğu söylenebilir. Bu durum, bir sinyalin veya gürültünün cevabı ve sunumu arasındaki ilişkiyi azaltmaktadır.



Şekil 1. İki Yanıtlı ve Lojistik Temelli Dağılımlar İçin Sinyal Algılama Teorisi

Şekil 1. SAK'ın iki temel fikrini ve parametrelerini göstermektedir. Birinci fikir, bir sinyal ya da gürültü gibi bir olayın sunumunun etkisinin, şekilde gösterilen lojistik dağılımlar gibi temel bir olasılık dağılımı ile temsil edilebileceğidir. Sinyal ve gürültü ile ilişkili dağılımların yalnızca konuma göre farklılık gösterdiği varsayılmaktadır. Altta yatan dağılımlar, uygulama alanına bağlı olarak birden fazla şekilde oluşturulabilmektedir. SAK'ın ikinci bir fikri, her denemede bir katılımcının bir yanıt kriteri kullanarak bir olayın meydana gelip gelmediğine karar vermesidir. Özellikle, katılımcı; sübjektif olay (Sansasyon gibi), ölçütün üzerine çıkarsa "evet" (bir sinyal bildirir) yanıtı verirken; aksi halde "hayır" cevabını vermektedir.

SAK belirsiz koşullarda bir uyarana ilişkin cevap örüntüsünü analiz etmeyi hedeflemektedir. Kısaca açıklamak gerekirse uyaranın var olduğu durum ile var olmadığı durum koşulları için verilen yanıtın doğruluğunu hesaplamaktadır.

Tablo 1. Uyaran ve Tepki Durumlarına Göre SAK İsimlendirmesi

	Tepki yok	Tepki var
Uyaran var	Kayıp (miss)	Doğru (hit)
Uyaran yok	Doğru kayıp (Correct Rejection)	Yanlış Alarm (False Alarm)

Tablo1.'de görüldüğü gibi olmayan uyarana gösterilen tepki yanlış alarm ve olan uyarana gösterilmeyen tepki kayıp olarak belirlenmektedir. Bu durumda hit/miss oranı üzerinden SAK parametreleri hesaplanmaktadır.

Belirli bir açık uçlu soruya verilen yanıt için gerçek kategori doğrudan gözlemlenememekte veya bilinmemektedir. Gerçek kategori örtüktür ve örtük kategorik değişken burada η olarak gösterilmektedir. Hakemlerin görevi sinyal algılamadır yani belirli bir açık uçlu maddeye verilen yanıt için, hakem, yanıtın hangi kategoriye ait olduğunu belirlemek zorundadır. HPM-SAK model ise SAK'ın bu yapısını hiyerarşik puanlama yaklaşımı ile modelin içine dâhil ederek hakem yargılarına ilişkin düzeltmelerin yetenek puanlamalarınaa etkisini belirlemektir.

SAK'ın temel parametreleri c ve d'dir. Parametre c, tepki kriterinin gürültü dağılım modundan uzaklığıdır (Macmillan & Creelman, 1991). Parametre d, ses veya gürültüye dair iki temel dağılım arasındaki mesafenin bir ölçüsüdür;

$$d = \frac{\psi_s - \psi_n}{\tau}$$

Burada, ψ_s ve ψ_n, sırasıyla, sinyal ve gürültü dağılımlarının modlarıdır ve τ, bir ölçek parametresidir. Lojistik dağılımlar için d basitçe, seçim kriteri parametresi α'nin logaritmasının iki katıdır, $d = 2\,log\,(\alpha)$ (Macmillan & Creelman, 1991; McNicol, 2005) ve " τ standart sapma σ ile ilişkilidir; $\sigma = \tau\pi/\sqrt{3}$. d'yi $\sqrt{3}/\pi$ ile çarpmak, meta-analizde kullanılan etki büyüklüğü ölçüsü (δ) olan standart sapmalardaki mesafeyi verir (Hasselblad & Hedges, 1995). Normal dağılımlar için, $\sigma = \tau$ ve d geleneksel d' ölçüsüdür. Bu ve diğer dağılımlara dayanan SAK modelleri ele alındığından, her ikisi de τ ile ölçeklendirilmiş olan mesafe ölçüsü için d ve yanıt kriteri için c kullanılarak notasyon birleştirilmiştir. SAK'nin önemli bir özelliği, d ve c parametrelerinin, sırasıyla, duyu (bellek, vb.) faktörlerini karar faktörlerinden ayırmasıdır. Şekil 1'de gösterilen sinyal dağılımı altında kriterin sağındaki alan, bir uyaran verildiğinde, "evet" cevabı verme ihtimalini yani bir isabet "hit" olasılığını vermektedir. Lojistik dağılım için bu durum aşağıdaki denklem ile verilir;

$$P(Y = 1|S) = \frac{1}{1 + \exp[(c - \psi_s)/\tau]}$$

Burada $P(Y=1|S)$, bir uyaranın verildiği göz önüne alındığında, katılımcının "evet" cevabı verdiği (bir sinyal bildirdiği) koşullu olasılığın ve $(c-\psi_s)/\tau$, kriterin sinyal dağılımı modundan ölçeklendirilmiş uzaklığıdır. Yukarıdaki denklemin sağ tarafı, lojistik dağılım için 'hayatta kalma fonksiyonu' dur; Alanın kriterin sağına geldiğini ve sadece kümülatif dağılım fonksiyonundan 1 eksik olduğunu göstermektedir. Benzer şekilde, gürültünün verildiği yani uyaranın verilmediği göz önüne alındığında katılımcının "evet" demesi olasılığı, yanlış bir alarmın olasılığı olan $P(Y=1|N)$, gürültü dağılımı altında kriterin sağındaki alan tarafından verilmektedir, bu durum aşağıdaki denklemle ifade edilir;

$$P(Y=1|N)=\frac{1}{1+\exp[(c-\psi_n)/\tau]}$$

Log, doğal logaritma olmak üzere, $logit(p)=\log(\frac{p}{1-p})$, logit dönüşümü, yukarıda gösterilen isabet ve yanlış alarm olasılıklarına uygulanırsa;

$$logit\ p(Y=1|S)=\frac{\psi_s-c}{\tau},$$

$$logit\ p(Y=1|N)=\frac{\psi_n-c}{\tau} \qquad (1)$$

Eşitlikleri elde edilir. Yukarıdaki logitler sırasıyla katılımcının bir sinyal verildiğinde "evet" dediği log olasılıklarını ve gürültüye yani sinyal verilmediğinde "evet" dediği log olasılığını temsil etmektedir. Denklem (1), logit dönüşümlü olasılıkların teorik parametrelerle basit bir ilişkiye sahip olduğunu göstermektedir. Örneğin, d ve c parametreleri için kolayca çözülebilir: logit hitleri ve logit yanlış alarmlar arasındaki fark d yi verir ve logit yanlış alarmların -1 katı c'yi verir; Olasılıklar için gözlemlenen orantılar parametrelerin tahminlerini verir (ayrıca, iki logitin toplamının $-1/2$ katı bir başka kriter ölçüsü c' yü verir).

1.2.2.1. Çoklu Gözlemci İle Eşit Olmayan Varyanslı SAK

SAK'ın temel fikri, gözlemcilerin yanıt kriterlerini algılamalarıyla birlikte tepki oluşturmasıdır. Yanıt ölçütlerinin kullanılması, SAK'ın bir bileşenidir; bu 'karar kuralı' şu şekildedir;

Eğer $c_{j(k-1)} < \psi_j < c_{jk}$ ise, $Y_j = k,$ (2)

Burada Y_j, 1'den K_j'ye kadar değişen aralıklarla ayrık değer k ile, j inci gözlemci için bir derecelendirme yanıtıdır, burada K_j, tepki kategorilerinin sayısıdır (genelde gözlemcilerle aynıdır), ψ_j, gözlemcinin algısını temsil eden örtük ve sürekli bir rastgele değişkendir ve c_{jk}, $c_{j0} = -\infty$ ve $c_{jK} = \infty$ ile kesin sıralı, $c_{j1} < c_{j2} < \ldots < c_{j(K-1)}$ olan j inci gözlemci için k ıncı kriterdir. Varyasyonun maddeler üzerinde olduğu j inci gözlemci için Y_j ve ψ_j'nin rastgele değişkenlerdir.

SAK'ın ikinci bir bileşeni, gözlemcilerin algıları ile gözlemlenen (veya gözlemlenmeyen) olaylar arasındaki ilişkiye ilişkin *algısal* ya da daha genel olarak *yapısal modeldir* (DeCarlo, 2010). Geleneksel eşit olmayan varyanslı SAK model şu şekildedir;

$$\psi_j = d_j x + \sigma_j^x \varepsilon_j, \quad (3)$$

Burada, sırasıyla, gürültü veya sinyal için $x = 0$ veya 1 (yani, gürültü dağılımının ortalaması sıfır noktası olarak kullanılır; maddeler tüm gözlemcilerde aynı olduğu için bir alt simge j'ye gerek yoktur). d_j, sinyal dağılımının ortalamasından j gözlemcisinin gürültü dağılım ortalamasına olan uzaklığı (Hata varyansının karekökü göre ölçeklendirilmiş dağılımda), σ_j ise bir ölçek parametresidir (sinyal varyansının gürültü varyansından farkıdır). ε_j, j inci gözlemcinin algılamasındaki rastgele değişimdir; $E(\varepsilon j) = 0$ ve varyans $V(\varepsilon_j)$, burada E beklenti operatörü ve V varyans operatörüdür, ε_j'nin ortalama sıfır olduğu kabul edilir; ε_j'nin x ile korele olmadığı da varsayılmaktadır. Tanımlama amacıyla, normal model için $V(\varepsilon_j)$, her bir gözlemci için bir bütünlük oluşturmaktadır.

Denklemlerin algı modelleri ve kararları (2) ve (3) ile birlikte eşit olmayan varyanslı SAK modelini verir. Özellikle,

$$p(Y_j \leq k|X = x) = p(\psi_j \leq c_{jk}|x) = p(d_j x + \sigma_j^x \varepsilon_j \leq c_{jk}) = p[\varepsilon_j \leq (c_{jk} - d_j x)/\sigma_j^x] .$$

Eğer $\varepsilon_j \sim N(0,1)$ ise o zaman,

$$p[\varepsilon_j \leq (c_{jk} - d_j x)/\sigma_j^x] = \Phi((c_{jk} - d_j x)/\sigma_j^x)$$

Ve böylece;

$$p(Y_j \leq k|x) = \Phi((c_{jk} - d_j x)/\sigma_j^x)$$

olur. Ve bu eşit olmayan varyanslı SAK Modelidir (DeCarlo, 2003).

SAK'ın temel yönleri, koşullu olasılıklar ve gizli değişken ψ 'nin varyansları ile gösterilebilir. Örneğin, Denklem (3) 'den gözlemcinin koşullu olasılık ve varyansları (maddeler üzerinde) aşağıdaki gibidir:

$$\mathrm{E}(\psi_j \mid x = 0) = \mathrm{E}(\varepsilon_j) = 0 \qquad \mathrm{V}(\psi_j \mid x = 0) = \mathrm{V}(\varepsilon_j)$$

$$\mathrm{E}(\psi_j \mid x = 1) = d_j \qquad \mathrm{V}(\psi_j \mid x = 1) = \sigma_j^2 \mathrm{V}(\varepsilon_j)$$

Burada d_j (ve ölçüt c_{jk}), $\mathrm{V}(\varepsilon_j)$ 'ye göre ölçeklendirilmektedir ve yukarıda belirtildiği gibi, $\mathrm{V}(\varepsilon_j)$, normal model için bütünlük oluşturmaktadır. Yukarıdaki sonuç, belirli bir gözlemci j için, sırasıyla $\sigma_j < 1$, $\sigma_j = 1$ veya $\sigma_j > 1$ ise, sinyal varyansı gürültü varyansından küçük, eşit veya büyük olabileceğini göstermektedir. Ayrıca Denklem (3)'te, x şartlı, ψ, ε'nin doğrusal bir dönüşümü ve normal ε normal $\psi \mid x$ dir. Eşit olmayan varyanslı SAK modelinin bireysel gözlemcilerin verilerine uyması d_j ve σ_j tahminlerini ve c_{jk} kriterleri tahminlerini verir.

Burada incelenen durum için, gözlemciler arasında ortak olan maddelerin potansiyel bir etkisi vardır ve bu nedenle koşullu kovaryanslar da göz önüne alınmalıdır. Örneğin, Denklem (3) gözlemcilerin j ve j' koşullu kovaryansı, gürültü için,

$\mathrm{Cov}(\psi_j, \psi_{j'} \mid x = 0) = \mathrm{Cov}(\varepsilon_j, \varepsilon_{j'}) = 0,$

$\mathrm{Cov}(\varepsilon_j, \varepsilon_{j'}) = 0$ gözlemciler arasında bağımsızlık varsayımından gelir. Benzer şekilde, sinyal için,

$$\mathrm{Cov}(\psi_j, \psi_{j'} \mid x = 1) = \mathrm{Cov}(d_j + \sigma_j\varepsilon_j, d_{j'} + \sigma_{j'}\varepsilon_{j'}) = \sigma_j\sigma_{j'}\,\mathrm{Cov}(\varepsilon_j, \varepsilon_{j'}) = 0$$

d_j, $d_{j'}$, σ_j ve $\sigma_{j'}$'nin belirli bir gözlemci için sabitler (maddeler üzerinde) olduğu göz önüne alındığında, yukarıdaki bilgiler, bir sinyalin varlığı veya yokluğuna bağlı koşulların, geleneksel eşit olmayan varyanslı SAK modelinde gözlemciler arasında algılamalar (ve yanıtlar) anlamında korelasyon olmadığını göstermektedir. Yukarıdaki değerlerin hata kovaryansını, yani sinyal veya gürültünün sunumu ile hesaplanmayan gözlemciler arasındaki kovaryansı temsil etmektedir. Madde etkisinin olasılığı, sıfır kovaryans varsayımının ihlal edilmesi anlamına gelmektedir.

1.2.3. HPM Model

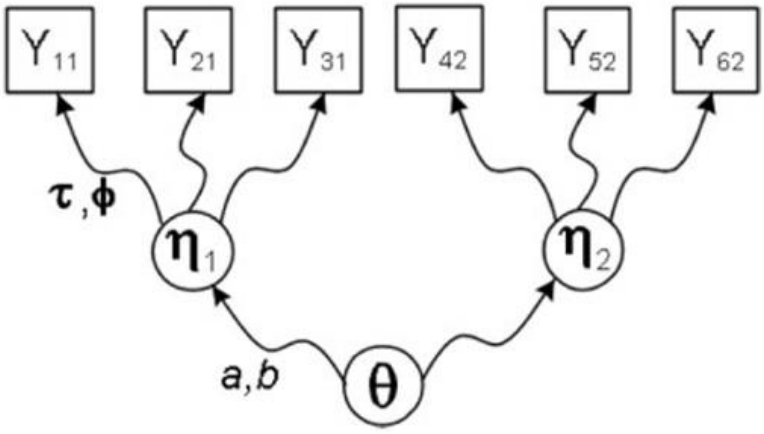
Patz (1996) ve Patz & diğerleri (2002); HPM'nin, YY puanlamasında hakemlerin kullanılmasının hiyerarşik bir veri yapısına yol açtığını kabul etmektedir. HPM'de hakemler tarafından sağlanan puanlar Rasch Modeli ya da diğer hakem puanlarının MTK yaklaşımlarında olduğu gibi, sınava giren kişinin yeteneklerinin doğrudan göstergesi değildir. Bunun yerine, puanlar, her bir açık uçlu maddeye verilen yanıtın gerçek veya "ideal" kategorisinin göstergeleridir; burada gerçek kategori, puanlama anahtarı tarafından tanımlanmaktadır.

Örneğin, SAK'de açık uçlu maddeye verilen yanıtları puanlayan hakemler, 1 = çok az ustalık veya hiç ustalık, 2 = az ustalık, 3 = geliştirilebilir ustalık, 4 = yeterli ustalık, 5 = makul ölçüde tutarlı ustalık ve 6 = net ve tutarlı ustalık, kategorilerin kullanılarak puanlama yapma eğitimi almıştır. Açık uçlu maddeye yanıt vermeyen sınava giren kişiler için sıfır kullanılmaktadır. Kategorilerin ayrıntılı açıklamaları da sağlanmaktadır (bz. Http://professionals.collegeboard.com/testing/satreasoning/scores/essay/guide).

Daha önce değinildiği gibi açık uçlu soruya verilen yanıt için gerçek kategori doğrudan gözlenememekte veya bilinememektedir. Gerçek kategori örtüktür ve örtük kategorik

değişken burada η olarak gösterilmektedir. Hakemlerin görevi sinyal algılamadır yani belirli bir açık uçlu maddeye verilen yanıt için, hakem, yanıtın hangi kategoriye ait olduğunu belirlemektedir. Bu durumda SAK, hiyerarşik yapının ilk aşamasındaki puanlayıcı modeli için uygun bir yaklaşım olmaktadır. Sinyal algılama yaklaşımı, hakemlerin kararlarının gerçek kategorinin yanıltıcı olabilen göstergeleri olduğunu kabul etmektedir. Aynı zamanda, hakemlerin gerçek kategorileri algılama yeteneğinin SAK yaklaşımı ile belirlenebileceğini savunmaktadır.

HPM'nin ikinci kademesinde, örtük kategoriler, bir MTK modeli aracılığıyla sınava giren kişinin yeterliğini gösteren sıralı göstergeler olarak görev yapmaktadır (durum, göstergelerin gözlemlenebilir olması yerine örtük olması sebebiyle normal MTK modelinden farklıdır). Bu nedenle, sınava giren kişi tarafından açık uçlu maddeye verilen yanıt, puanlama anahtarı tarafından tanımlanan kategorilerden birine ait olarak görünmektedir ve açık uçlu maddeye verilen yanıtın ait olduğu (doğru) kategori, sınava giren kişinin yeterliliğinin bir göstergesidir. Burada kullanılan gösterimde, gözlemlenen hakem puanları Y, açık uçlu maddeye verilen yanıtın gerçek kategorisi η'nın göstergeleridir ve gerçek kategori η, sırasıyla, sınava giren kişilerin yeterliklerinin θ(tetha)'nın bir göstergesidir.



Şekil 2. Patz (1996) ve Patz ve Diğerlerinin(2002) HPM Modeli

Şekil 2. her bir sınava giren kişinin iki maddeye (Açık uçlu maddeler, YY maddeleri vb.) yanıt verdiği ve her yanıtın üç hakem tarafından puanlandığı bir durum için HPM temsilini göstermektedir. Altı hakemin puanlarına sıralı tepkiler (örn., 1 ila 6 skor)

gözlemlenmekte ve j inci hakem ve l inci madde için Y_{jl} olarak Şekil 2.'de gösterilmektedir. Hakemler, birinci madde için örtük kategorik değişken η_1 ve ikinci madde için η_2 ile gösterilen, her yanıt için gerçek gizli kategoriyi bulmaya çalışmaktadır. η_l'den Y_{jl}'ye doğru olan okların, ilişkinin doğrusal olmadığını belirtmek için kavisli olduğu ve özellikle Y_{jl}'in olasılığının doğrusal olmayan bir fonksiyonla η_l'e bağlı olduğu Şekil 2.'de görülmektedir.

Temel Düzey 1 (sinyal algılama) parametreleri Şekil 2.'de gösterildiği gibi τ_{jl} ve ϕ_{jl}'dir ve bunlar sırasıyla hakem hassasiyeti ve hakem katılığı parametreleridir. HPM'nin ikinci düzeyinde, açık uçlu maddeye verilen yanıtın gerçek kategorisi, η_l, bir sınava giren kişinin yeterlik derecesi θ'nın sıralı göstergesi olarak kullanılmaktadır. Burada oklar yine örtük kategorik değişken η_l'nin (yani olasılıklarının), bir MTK modeli (burada Genelleştirilmiş Kısmi Puanlama Modeli kullanılır, diğer MTK modelleri de kullanılabilir) aracılığıyla sınava giren kişilerin yeterlikleri olan θ ile doğrusal olmayan bir ilişkiye sahip olduklarını göstermek için kavislidir. Düzey 2 parametreleri sırasıyla normal ayırt edicilik ve madde güçlüğü parametreleri olan a_l ve b_{lm}'dir.

Yukarıda belirtildiği gibi, HPM, hakemin puan vermeye yönelik bir MTK yaklaşımı kullanıldığında ortaya çıkan soruna hitap etmektedir (diğer bir deyişle, hakem sayısının arttırılması, gittikçe artan düzeyde yeterlik tahminlerini vermektedir). Özellikle, bir MTK yaklaşımında daha fazla hakem, θ hakkında doğrudan daha fazla bilgi sağlarken, HPM için Şekil2.'de gösterildiği gibi, daha fazla hakem kullanmak η_l hakkında daha fazla bilgi sağlarken, doğrudan θ hakkında bir bilgi sağlamamaktadır.

Bir test, birden fazla hakem yargılarına dayalı puanlama yapılan madde başka bir problem ortaya çıkmaktadır. Şekil 2.'de gösterildiği gibi, ilk üç hakem birinci maddeye, ikinci üç hakem ise ikinci maddeye yuvalanmaktadır. Facets Modeli (Linacre, 1989) gibi hakem verileri için yaygın olarak kullanılan modeldeki altı hakemden alınan puanlar kullanılırsa, hakemlerin maddeler arasında yuvalanmış olması gerçeği göz ardı edilebilmektedir. Yuvalama, örneğin, 1, 2 ve 3 hakemlerinden alınan puanlar kısmen koraledir çünkü üç puanlayıcının hepsi aynı maddeyi (birinci maddeyi) puanlamaktadır. Aynı durum, ikinci maddeyi puanlayan 4., 5. ve 6. hakemlerinin hepsi için de geçerlidir. Yuvalanma

nedeniyle ortaya çıkan korelasyonun yok sayılması, Patz ve diğerlerinin (Richard J Patz et al., 2002), Donoghue & Hombo (2000) ve Wilson & Hoskens (2001) in de not ettiği gibi, negatif yanlılık yeterliliğin standart hatalarının tahminlerini vermektedir. Öte yandan, HPM, yuvalanmayı tanımakta ve negatif yanlılık önyargıyı düzeltmektedir.

1.2.3.1. Patz Puanlayıcı Modeli

HPM'nin ilk düzeyi için Patz ve arkadaşları (2002); ayrıca Mariano ve Junker (2007) tarafından kullanılan sinyal algılama benzeri model şu şekilde yazılabilir:

$$p(Y_{jl} = k \,|\eta_l = \eta) \propto \exp\left\{-\frac{1}{2\psi_{jl}^2}[k - (\eta - \phi_{jl})]^2\right\}, \qquad (4)$$

Burada Y_{jl}, j hakeminin l maddesine verdiği yanıttır ve K yanıt kategorileri ile ayrık bir k puanı oluşmaktadır. (Bu şart olmasa da, yanıt kategorilerinin sayısı, genel olarak uygulamada olduğu gibi, farklı hakemler ve maddeler arasında aynı olduğu varsayılmıştır) , η_l, l inci madde için bir örtük kategorik değişken, ψ_{jl}^2, j hakemi için (ve l maddesi) bir varyans parametresidir ve ϕ_{jl}, bir hakem katılık parametresidir. Diğer bir deyişle, daha yüksek değerler daha sert bir hakemin göstergesidir.

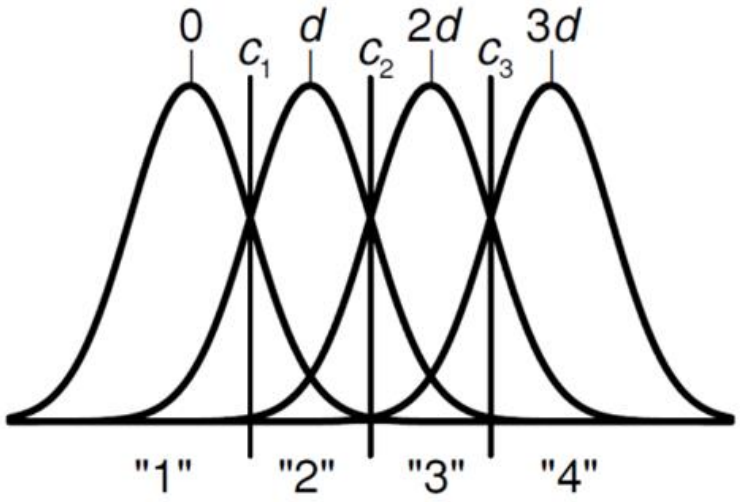
Eşitlik (4), her yanıt kategorisinin olasılıkların neredeyse normal dağıldığı sinyal algılama benzeri bir modeldir. Patz ve diğerlerinin (2002) belirttiği gibi, ψ_{jl}^2, bir hakemin güvenilirliğe dair eksikliğinin bir ölçüsüdür; diğer taraftan, $\tau_{jl} = 1/2\psi_{jl}^2$, hakem hassasiyetinin bir ölçüsüdür. Diğer bir hakem parametresi, ϕ_{jl}, hakemin sert(pozitif değerler) veya cömert(negatif değerler) olup olmadığını gösteren bir katılık parametresidir; sırasıyla hakem düşük veya yüksek puanlar verme eğilimindedir. Patz ve diğerleri (2002) tarafından kabul edilen (4) denklemi ile ilgili bir problem, "en güvenilir hakemler" (yani, küçük ψ_{jl}^2 değerlerine sahip olanlar) "en düşük hakem önyargı(bias) parametreleri"ne sahip olanlar olması eğilimi göstermesidir. Bunun nedeni, ψ_{jl} nispeten küçük olduğunda, bir fonksiyon olarak ϕ_{jl}'nin olasılığı $(-0.5, 0.5)$ Aralığı boyunca neredeyse sabit ve bu aralığın dışında sıfıra yakın olmasıdır (Çünkü gerçek kategori dışındaki yanıt kategorisindeki bir skorun olasılığı sıfıra yakındır). ϕ_{jl} için olasılık -0.5

ila 0.5 arasında hemen hemen aynı olduğundan, o aralıktaki ϕ_{jl} için benzersiz bir değer belirlemek zordur. Patz ve diğerleri (2002) tarafından tespit edildiği gibi, ϕ_{jl}'nin posterior dağılımı neredeyse sabit olduğu için (ψ_{jl} = 0.05 ve 0.06 tahminleri ile) iki güvenilir hakem için ϕ_{jl} nin benzersiz bir değerini belirlemekte problemler oluşmaktadır.

(4) Denkleminin bir diğer kısıtlaması, sertlik veya cömertlik dışında hakem etkilerini yakalayamamasıdır. Bununla birlikte, merkezi eğilim veya ranj daralması (Myford & Wolfe, 2004) gibi geniş kapsamlı değerlendirmelerde görünen çeşitli hakem etkileri de bulunmaktadır. Geleneksel SAK modeline dayanan örtük sınıf modeli, bu tür etkilerle kolayca başa çıkabilmektedir.

1.2.3.2. Düzey 1: Hakem Tutumuna Dayalı Örtük Sınıf SAK Modeli

Hakem yargılarına dayalı maddelerin puanlamasına karışan psikolojik süreçlerin SAK (Green & Swets, 1988; Neil A Macmillan & Creelman, 2005; Wickens, 2002) çerçevesinde yararlı bir şekilde anlaşılabileceği daha önce öne sürülmüştür (Decarlo, 2002, 2005) ve başarıyla psikoloji ve tıpta kullanılmıştır. Bu uygulamalar SAK'ın örtük sınıf genişletmesini içerir (Decarlo, 2002).



Şekil 3. Hakemlerin Çoklu Puanlamaları İçin SAK Modeli.

Hakem yargılarına dayalı puanlanan maddeler, SAK'ta iki temel açıdan kavramsallaştırılmıştır: 1.Bir açık uçlu maddeye verilen yanıtın kalitesinin hakem

tarafından algılanışı ve 2.hakemin karar kriterlerini kullanması. Şekil3. Bu modelin temel yapısını göstermektedir. Hakemlerin, dört örtük kategoriyi tespit etmek için 1 ila 4 arasındaki tepkileri kullandıkları varsayılmaktadır. SAK'taki temel bir fikir, bir hakemin kararının, ψ örtük sürekli rastgele değişken olmak üzere, açık uçlu maddeye verilen yanıtın kalitesine ilişkin kendi algısına (ψ) dayandırılmasıdır. Algıların, normal ya da lojistik gibi bir olasılık dağılımından oluştuğu varsayılmaktadır (DeCarlo, 1998).

Şekil 3.'te gösterildiği gibi, ψ 'nin dağılımı, örtük kategorilerin her biri için farklı bir konuma sahiptir ve dört kategori için dört konum oluşmaktadır. Örneğin, birinci kategoriden bir yanıt için, yanıt kalitesinin hakem algısı (ψ) ilk olasılık dağılımından; ikinci kategoriden bir yanıt için, hakem algılaması ikinci olasılık dağılımından elde edilmektedir.

Şekil 3.'te olduğu gibi, algılamaya dayalı dağılımlar arasındaki uzaklıklar olan d, bir hakemin örtük kategorileri algılama (ayırt etme) yeteneğini yansıtmaktadır. Dolayısıyla, d, aynı geleneksel SAK'taki gibi, hakem algısının veya hakem hassasiyetinin bir ölçüsünü sağlamaktadır. Yaklaşımın, örtük sınıf modellerinin parametrik olmayan ölçüm güvenilirliklerini tanımlamak için Clogg ve Manning (1996) tarafından kullanılan yaklaşımla aynıdır (DeCarlo, 2002, 2005).

Şekil 3. ayrıca basitleştirici bir varsayımı göstermektedir; bu algılama dağılımları arasındaki uzaklık d'nin kategoriler arasında aynı olduğudur. DeCarlo (2002) bunu *eşit mesafe modeli* olarak adlandırılmıştır. Bir de eşit *olmayan mesafe modeli* durumu vardır, ancak eşit mesafe modeli daha cimridir ve daha önceki araştırmalar uyum indekslerinin sınırsız model üzerinde eşit mesafe modelini seçme eğiliminde olduğunu bulmuştur (DeCarlo, 2002, 2005). Bu varsayımı rahatlatıcı etkiler de güncel araştırmalarda incelenmektedir.

SAK'ın ikinci bir temel fikri, hakemlerin açık uçlu maddeye verdiği yanıtın kalitesi hakkında, *algılarını*, karar alanını dört yanıt kategorisine bölen *yanıt kriterleriyle* birlikte kullanarak, bir karara varmalarıdır; Üç kriter Şekil 3.'te dikey çizgiler olarak gösterilmiştir. Böylece, belirli bir yanıtın hakem algısı ilk ölçüt c_1'in altındaysa, hakem "1" yanıtını verirken; birinci ve ikinci kriterler arasında ise, hakem "2" yanıtını

vermektedir. Yorumun SAK açısından yararlı bir yönü, elde edilen kriterlerin karşılaştırılabileceği referans noktaları önermesidir. Örneğin, referans noktaları, doğru sınıflandırmaları en üst düzeye çıkarmaya çalışmak, olasılık oranları üzerine dayanan kararlar vermek vb. gibi çeşitli karar özelliklerini yansıtmaktadır (Egan, 1975). Şekil 3, bitişik dağılımların kesişme noktalarında bulunan kriterleri (örtük kategorilerin sayısı tepki kategorilerinin sayısına eşit olduğunda ve bitişik dağılımların olasılık oranlarının tek olduğu durumda geçerlidir) göstermektedir. Bu, SAK'ın önceki tartışmalarla (Egan, 1975; Wickens, 2002) ve büyük ölçekli değerlendirmelerdeki ölçüt yerleri hakkındaki son araştırmalarda (Decarlo, 2008) bulunan sonuçlarıyla ilgilidir.

Örtük sınıf sinyal algılama modeli Şekil 3. ve yukarıda tartışılan varsayımları izler ve şu şekilde yazılabilir:

$$p(Y_{jl} \leq k \,|\eta_l = \eta) = \mathrm{F}(c_{jkl} - d_{jl}\,\eta_l)\,, \qquad (5)$$

Burada Y_{jl}, j inci hakemin l inci maddeye verdiği yanıt ve K yanıt kategorileri ile ayrık bir puan olduğu yerde k; η, 0 ile $M-1$ arasında olmak üzere, η_l, M üzerinde değerler alan l inci madde için örtük kategorik değişken, (değerleri d_{jl} üzerinde eşit mesafe kısıtlamasını sağlar, Şekil 3.), F bir kümülatif dağılım fonksiyonudur (lojistik veya normal gibi dağılımların bir konum ailesi için), d_{jl}, j inci hakem ve l inci madde için bir algılama parametresidir, c_{jkl}; $c_{j1l} < c_{j2l} < c_{j3l} < \cdots < c_{j(K-1)l}$ ayrıca $c_{j0l} = -\infty$ ve $c_{jKl} = \infty$ olmak üzere, j inci hakem, l inci madde ve k ıncı yanıt kategorisi için $K-1$ kesin sıralı yanıt kriteridir. Lojistik model için, d_{jl} ve c_{jkl} lojistik dağılımın varyansının kareköküne göre ölçeklenmiştir, $\pi^2/3$.

Kategoriye özgü yanıt kriterlerine izin vererek, (5) denkleminin örtük sınıf SAK modeli çeşitli hakem etkilerini ele alabilmektedir. Hakem etkileri, söz konusu hakemlerin gözlemlenen nihai kategorileri kullanmamak, ranj daralması vb. sert ve cömert olma eğilimleridir. Örneğin, merkezi eğilim (Myford & Wolfe, 2004) bazı araştırmacıların, son kategorileri kullanmama (örneğin 1'den 6'ya kadar olan ölçekte 1 ve 6 kullanmamaya eğilimli olması gibi) eğiliminde oldukları gözlenmektedir.

SAK'ta, bu etki, hakemler, en yüksek kriteri en sağa, en düşük kriteri ise en sola yerleştirirlerse (ve böylece, son kategorilerin kullanılma olasılıkları küçüktür) ortaya çıkmaktadır; Bu patern büyük çaplı bir değerlendirmenin örtük sınıf SAK analizinde son zamanlarda bulunmuştur (Decarlo, 2008).

Patz ve diğerlerinin (2002) puanlayıcı modelinde merkezi eğilim ϕ_{jl} tarafından basit bir şekilde yansıtılamamaktadır, çünkü ϕ_{jl} yalnızca genel katılık düzeyindeki değişikliklere izin verirken, yanıt kategorileri arasındaki farklı katılık derecelerine vermemektedir.

Başka bir örnek olarak, hakemler bazen, tüm yanıt kategorilerini kullanmayabilirler, (1-6 arası bir ölçekte yalnızca 2 ila 6 arası yanıtlar vermiş olduklarını durumda) bu ranj daralmasını(aralığın kısıtlanması) göstermektedir.

Yine SAK'ta bu tür bir etki, hakemin birinci kategori (c_{j1l}) için düşük bir kriteri olduğu ve dolayısıyla nadiren (veya asla) "1" yanıtını verdiği anlamına gelir. Bununla birlikte, yanıt kategorilerinden yalnızca biri için katılığın var olabilmesine karşın bu, Patz ve diğerlerinin (2002) modelinde daha yüksek (veya daha düşük) genel katılık düzeyi (ϕ_{jl} ile ölçülen) gibi görünmektedir.

1.2.3.3. Düzey 2: Örtük Göstergeli Madde Tepki Kuramı

HPM'nin ikinci düzeyi, her bir maddenin örtük kategorik değişkeni (diğer bir deyişle, gerçek kategoriler), sınava giren kişinin yeterliliğinin sıralı göstergeleri olarak ele alınmaktadır.

Örneğin, Patz ve arkadaşları (2002) Kısmi Puanlama Modelini (Masters, 1982), buna karşın DeCarlo (2011) Genelleştirilmiş Kısmi Puanlama Modeli kullanmıştır (Muraki, 1992).

Her iki model de özellikle bitişik kategori logitelerini kullanmaktadır (Agresti, 2002), Genelleştirilmiş Kısmi Puanlama Modeli şu şekilde yazılabilir:

$$\log\left[\frac{p(\eta_l = \eta + 1|\theta)}{p(\eta_l = \eta|\theta)}\right] = a_l\theta - b_{lm} \qquad (6)$$

Burada η_l, η değeri 0'dan $M - 1$'e kadar çıkaran l maddesi için örtük kategorik bir değişkendir. Örtük sınıf sayısının M olduğu kabul edilir; puanlama anahtarında verilen yanıt kategorilerinin sayısı ile aynıdır. K, yukarıda bahsedilmiştir, ancak bunun varsayılmasına gerek yoktur. θ, (sınava giren kişinin yeterliliği) $N(0,1)$ olarak varsayılan bir sürekli örtük değişkendir. a_l, l inci madde için bir madde ayırt edicilik parametresidir ve $b_{lm}, m = \eta + 1$ olan $M - 1$ "madde adım" parametresidir (Masters, 1982). Böylece adım parametreleri b_{l1}, b_{l2} vb. şeklindedir. Adım parametreleri bazen geçiş parametreleri olarak da adlandırılmaktadır (De Ayala, 2013).

Eşitlik (6), $\eta = 0$ için Kategori0 a karşı Kategori1 in (örn, η + 1), $\eta = 1$ için Kategori 1 e karşı Kategori 2 nin vb. yanıt olasılığının oranının logunu modeller.

Yukarıdaki diğer dönüşümleri (bitişik kategori logitleri yerine kümülatif logitler gibi) kullanmak, dereceli yanıt modeli (graded response model); gibi diğer MTK modellerini vermektedir (Samejima, 1969). Kısmi Puanlama Modeli, a_l, maddeler arasında eşit olacak şekilde ayarlandıysa, (6) 'ya göre izah edilir. Masters (1982) tarafından "adım" veya bitişik puanlar arasındaki geçiş olasılıkları olarak motive edildiği için, genelleştirilmiş (ve kısmi) kredi modelinde bitişik kategori logiteleri kullanılmaktadır. Örneğin, birinci adım parametresi b_{l1}, l maddesi için, sıfır puanından bir puanına çıkma ihtimalini belirler; Ikinci adım parametresi b_{l2}, bir puandan iki puanına çıkma ihtimalini verir, vb. Bu açıkça (6) ile gösterilmiştir. Model ayrıca sıklıkla olasılıklar olarak yazılmaktadır; bu durumda (6) şu şekilde yeniden yazılabilir:

$$p(\eta_l = \eta|\theta) = \frac{e^{\sum_{m=0}^{\eta} a_l\theta - b_{lm}}}{\sum_{v=0}^{M-1} e^{\sum_{g=0}^{v} a_l\theta - b_{lm}}} ,$$

Burada $\sum_{\boldsymbol{m=0}}^{\boldsymbol{0}}(a_l\theta - b_{lm}) \equiv 0$ dır(Masters, 1982).

Marjinal örtük sınıf boylarının tahminleri (aşağıda verilmiştir), $p(\eta_l)$, $p(\eta_l \mid \theta)$ çarpımının hesaplanmasıyla ve her bir kuadratür noktası θ_q'da düğüm ağırlıkları w_q' iken tüm düğümlerin toplanmasıyla elde edilebilir.

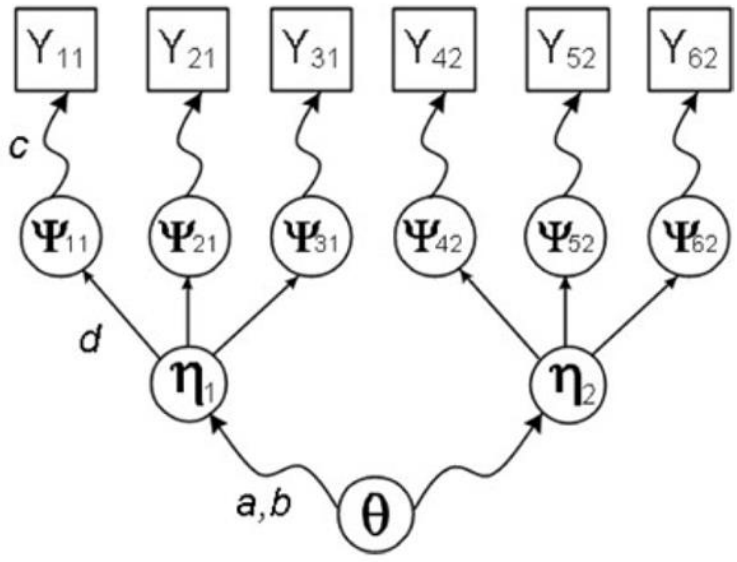
1.2.4. HPM-SAK Model

Şekil 4. birinci düzey modeli olarak örtük sınıf SAK modeli ve ikinci düzey modeli olarak bir MTK modeli bulunan tam HPM-SAK modelini göstermektedir. Daha önce olduğu gibi, kavisli oklar doğrusal olmayan ilişkileri belirtmek için kullanılmaktadır.

İlk düzeyde, hakem yanıtları Y_{jl} (aslında tepki olasılıkları), Şekil 3.'te gösterildiği gibi, yanıt olasılıklarının hakem kriterlerinin konumları olan c_{jkl} ye bağlı olmasıyla, açık uçlu maddeye verilen yanıtın kalitesi, hakem algıları ψ_{jl} ile doğrusal olmayan bir ilişkiye sahiptir.

Hakem algısının (ψ_{jl}) konumu, sırasıyla açık uçlu maddeye verilen yanıtın gerçek kategorisine, η_l ye bağlıdır; Şekil 3'teki η_l'den ψ_{jl}'ye düz oklar doğrusal bir ilişkiyi gösterir ve özellikle, örtük kategori η_l bir birim artıkça, ψ_{jl} nin ortalaması d_{jl} tarafından kaydırılmaktadır.

Şekil 3.'te gösterildiği gibi, d_{jl}, hakemin algısal dağılımları arasındaki mesafeyi belirtmektedir ve hakem hassasiyetinin bir ölçüsüdür. İkinci düzeyde, gerçek kategoriler η_l, Şekil 4.'te gösterildiği gibi, ayırt edicilik ve madde adımı parametreleri a_l ve b_{lm} ' ile sınava giren kişinin yeterliliği olan θ nın göstergesi olarak görev yapmaktadır.



Şekil 4. HPM-SAK Modeli

Şekil 4. HPM-SAK'ın, birinci düzey için örtük sınıf SAK modeli ve ikinci düzey için bir MTK modeli ile daha yüksek dereceli faktör modelinin bir türü (type of higher-order factor model) (Bollen, 1989) olduğunu göstermektedir. Örtük sınıf SAK modelinin ayrık faktör modelleri, ayrık MTK modelleri ve diğer modeller ile (örn. konumlandırılmış örtük sınıf modellerinin) ilişkileri (Decarlo, 2002, 2005, 2008) kaydedilmiştir. HPM'nin Rasch Modeli ile ilişkileri ve genellenebilirlik teorisi modelleri, Patz (1996) ve Patz & diğerleri (2002) tarafından tartışılmıştır.

Tam HPM-SAK modeli, yukarıda verilen Düzey 1 ve 2 bileşenlerini içermektedir. Y, sınava giren kişiler için yanıt değişkenlerinin vektörü dür. $Y = (Y_{11}, Y_{12}, \ldots Y_{1L}, Y_{21}, \ldots, Y_{2L}, \ldots Y_{J1}, \ldots, Y_{JL})$ şeklinde gösterilmektedir. Burada Y_{jl}, j inci hakem ve l inci madde için yanıt değişkenidir (sınava giren kişiye göre değişir). Uygulamada yaygın olarak ortaya çıkan durumlar için, bu matrisin bazı vektörleri eksik olabilmektedir. (Örneğin, hakemler yalnızca bazı sınava giren kişileri puanladığı ve aynı zamanda yalnızca maddelerin kimisini puanladığı durumlar gibi.) $\eta = (\eta_1, \eta_2, \ldots, \eta_L)$, her biri M kategori içeren YY maddelerinin L örtük kategorik değişkenlerini göstermek üzere HPM-SAK modeli aşağıdaki gibi ifade edilir;

$$p(Y) = \sum_{\eta} \int_{\theta} p(Y|\eta, \theta) p(\eta|\theta) p(\theta) d\theta \qquad (7)$$

Burada $p(Y|\eta, \theta)$, modelin hakem bileşenidir (birinci düzey) ve $p(\eta|\theta)$, YY maddelerinin bileşenidir (ikinci düzey). Düzey 1'de iki önemli varsayım yapılmıştır. Birincisi, örtük değişkenler η üzerinde koşullu olan gözlenen skorlar (Y), yetenekten (θ) bağımsızdır. İkincisi, η örtük değişkenlerin vektörü üzerinde koşullu, skorlar bağımsızdır. Bu iki varsayım modelin hakem bileşenini aşağıdaki gibi basitleştirir:

$$p(Y|\eta, \theta) = p(Y|\eta) = \prod_{jl} p\left(Y_{jl}|\eta\right), \qquad (8)$$

Burada j, hakem, l ise YY öğesini belirtir. Koşullu olasılıklar $p\left(Y_{jl}|\eta\right)$, kümülatif olasılıkların farklılaştırılmasıyla (5) denkleminden elde edilir. Düzey 2 için verilen θ nın L örtük değişkenlerin koşullu bağımsızlığı varsayımı yapılmaktadır.

$$p(\eta|\theta) = \prod_{l} p(\eta_l|\theta) \qquad (9)$$

Burada (6) 'dan koşullu olasılıklar $p(\eta_l|\theta)$ elde edilir ve olasılıklar açısından yukarıda gösterildiği gibi yeniden yazılır. (8) ve (9) 'nun (7)'de yerine konulması ve (5) 'in yanıt olasılıkları için değiştirilmiş formunun ve örtük sınıf olasılıkları için (6)'nin olasılık formunun kullanılması, tam HPM-SAK modelini verir.

1.2.4.1. HPM-SAK İçin Parametre İyileştirilmesi

Yukarıda açıklandığı gibi maksimum olasılık tahmini (MOT) veya Posterior mod tahminlerini (PMT) kullanarak HPM-SAK için parametrenin iyileştirilmesi daha önce birkaç simülasyon çalışmasında incelenmiştir (DeCarlo, 2008, 2010; DeCarlo & Kim, 2009; DeCarlo et al., 2011). Örneğin, tamamen çapraz tasarımların (diğer bir deyişle, tüm hakemlerin tüm açık uçlu maddeye verilen yanıtları puanladığı) simülasyonları için sonuçlar DeCarlo (2008, 2010) ve DeCarlo & Kim (2009) tarafından sunulmuştur.

Temel bulgular, SAK modeli (birinci düzey) için hakem parametrelerinin genel olarak başarılı bir şekilde iyileştirildiği şeklindedir. İkinci düzey (MTK modeli) için, yalnızca iki madde olduğu zaman Genelleştirilmiş Kısmi Puanlama Modeli için madde parametreleri zayıf bir şekilde iyileştirilmiş; üçüncü bir madde eklendiğinde madde parametrelerinin iyileştirilmesi büyük ölçüde geliştirilmiştir. Sınır sorunları da yaşanmıştır; Bununla birlikte, PMT'nin Bayes teklik sabitleri ile birlikte kullanılmasıyla iyi bir parametre iyileştirilmesinin sağlandığı gösterilmiştir.

DeCarlo & Kim (2009) tarafından, birçok gerçek uygulama değerlendirmesinde elde edildiği gibi (yani, her hakem sınava giren kişilerin yalnızca bir alt kümesini değerlendirdiği ve bu sebeple kayıp verilerin olduğu) tamamlanmamış tasarımlar için HPM-SAK'taki parametre iyileşmesini inceleyen simülasyonlar sunulmuştur. Kim, modelin SAK kısmı için, HPM-SAK'ın parametre iyileştirmesinin, uygulamada bulunan bir algılama aralığı (d) için iyi olduğunu bulmuştur (Örn., Yaklaşık 1 ila 6; bkz. (Decarlo, 2005, 2008a,2008b, 2010).

Madde parametrelerin tahmini genelleştirilmiş Kısmi Puanlama Modeli için sadece iki madde olduğunda marjinaldir, üçüncü bir madde eklendiğinde ise model daha da iyileşir. DeCarlo ve Kim (2009), HPM-SAK, 2. düzeydeki θ'nın doğrudan göstergeleri olarak

çoktan seçmeli maddeleri kullanarak genişletildiğinde, açık uçlu madde parametrelerinin (düzey 2) tahminleri, yalnızca bir veya iki açık uçlu madde ile olsa dahi, büyük ölçüde iyileşmiş olduğunu göstermiştir. Bir diğer ilginç sonuç ise, Düzey 1'de hakem parametrelerinin tahmininin, 2. düzeyde çoktan seçmeli maddeler eklendiğinde biraz daha iyileşmiş gibi görünmesidir. Özetle, simülasyon çalışmaları, HPM-SAK modeli için hakem parametrelerinin genel olarak iyileştirilebileceğini göstermiştir.

Madde parametreleri de uygun şekilde iyileştirilmiş gibi görünse de, Genelleştirilmiş Kısmi Puanlama Modeli, Düzey 2 modeli olarak kullanıldığında, ikiden fazla madde veya diğer ek bilgiler (yani çoktan seçmeli maddeler tarafından sağlananlar gibi) yeterli parametre iyileştirmesi için gerekli olduğu söylenmektedir.

1.3. Test Geliştirme

Test geliştirme psikometride en çok araştırılan konuların başında gelmektedir. Bu sürecin aşamaları konusunda Murphy & Davidshofer (1988), Downing & Haladyna (2006), Cronbach (1989) ve Gronlund (1982) gibi pek çok kişi, çeşitli araştırmalar yapmışlardır. Bu çalışmalar ışığında genel olarak test geliştirme süreci aşağıdaki basamaklarla özetlenebilir.

- Testin amacının belirlenmesi: Testin kimlere(hangi hedef kitleye) uygulanacağını ve test puanlarının hangi amaç için kullanılacağını belirleme aşamasıdır.
- Test ile ölçülecek özelliğin belirlenmesi: Başarı, tutum, kaygı, motivasyon vb. ölçülmesine odaklanılan kavramların belirginleştirildiği adımdır. Burada hangi kavram üzerinde çalışıldığı ve ne gibi kriterler ile değerlendirme yapılacağı netleştirilmektedir.
- Ölçülecek özelliğe uygun madde türlerinin belirlenmesi: Maddelerin yazılması, ölçülen özellik grubuna uygun madde formatlarının belirlenmesi ve madde havuzunun oluşturulması süreçlerini kapsamaktadır.

- Test maddelerinin teknik denetiminin yapılması: Maddelerin ölçülecek özelliği ölçmedeki yeterliğine yönelik teknik denetimlerin yapılması ve dil açısından anlaşılırlığının incelenmesi adımıdır.
- Uzman görüşleri alınarak test maddelerinin düzeltilmesi: Uzman görüşleri yardımı ile test maddelerinin düzenlenmesi, yönergenin hazırlanması, düzenlenen sorular ile ön değerlendirme formunun oluşturulması aşamasıdır. Bu aşamada mümkünse bu formun yönergesinin ve sorularının anlaşılırlığını yoklamak için küçük bir grup üzerinde uygulama yapılması tercih edilmektedir.
- Ön deneme formunun hazırlanması: Küçük bir grup üzerinde uygulanmış testin düzeltmelerinin yapıldığı adımdır. Bu adımda maddelerin varsa hataları ve yönergenin varsa anlaşılırlığı bozan kısımları düzeltilerek ön deneme formu oluşturulmaktadır.
- Ön uygulamanın yapılması: Düzeltmelerin sonrasında oluşturulan formun uygulamaya hazır hale getirilmesi ve belirlenen örneklem üzerinde uygulanması aşamasıdır. Kitapçıklar hazırlanır, yönergeler hazırlanır ve ön uygulama gerçekleştirilir.
- Ön uygulamadan elde edilen verilere dayalı test ve madde istatistiklerinin belirlenmesi: Uygulama sonrası testten elde edilen veriler ile test ve madde istatistiklerinin saptanması, madde analizlerinin yapılması, geçerlik ve güvenirlik analizlerinin de yapılması ile teste son şeklinin verilmesi basamağıdır.
- Uygun maddelerin tespiti ve uygulama için son formun hazırlanması: Analizler sonucunda uygulamaya uygun maddelerin seçimi düzeltmeleri ile nihai formun oluşturulması aşamasıdır. Burada istenene uygun madde istatistiklerine sahip maddeler seçilmektedir. İstenen şekilde çalışmayan maddelerde düzelmeler yapılmakta, varsa hatalar giderilmekte ve test son haline getirilmektedir.
- Testin uygulanması: Nihai formun belirlenen örneklem üzerinde uygulanması aşamasıdır. Bu aşamada ön uygulamanın yapılması adımları gerçekleştirilmekte ve test hedef kitle üzerinde uygulanmaktadır.

1.3.1. Madde Değerlendirme Yöntemleri

Madde değerlendirme aşaması testin kalitesini doğrudan etkileyen önemli bir aşamadır. Bu aşamada ön değerlendirme öncesi maddeler hakkında doğru değerlendirmelerde bulunmak sonraki her aşamanın daha sağlıklı ilerleyebilmesi için temel oluşturmaktadır. ETS, GRE, PISA ve TIMSS gibi sınavlarda da maddelerin geliştirilip pilot uygulamaya seçilmesi aşamalarında benzer yöntemler izlenmektedir.

Örneğin ETS; Çerçeve Oluşturma, Tasarı oluşturma, Soru yazımı, İçeriğin gözden geçirilmesi, Yanlılık denetimi, Yayın İncelemesi, Paydaş İncelemesi, Pilot Uygulama, İstatistiksel değerlendirme, Kullanıma hazır hale getirme gibi adımları izlemektedir. Çerçeve Oluşturma safhasında soru türleri seçilmekte ve sınavın nasıl uygulanacağı belirlenmektedir. Tasarı oluşturma aşamasında testteki her soru için ayrıntılı betimlemeler oluşturulmaktadır. Tasarılar bütün çalışmanın birbiri ile uyumlu olmasını sağlamaktadır. Soru yazımı aşamasında hazırlanan tasarı, test edileni öğretmede yıllarca tecrübesi olan soru yazarlarına yönlendirilmektedir. İçeriğin gözden geçirilmesi safhasında sorunun cevabının doğruluğunu belirlemek ve çeldiricilerden hiçbirinin sorunun herhangi olası yorumlanma durumunda doğru olmadığını kesinleştirmek için diğer test geliştiricileri tarafından gözden geçirilmektedir. Yanlılık denetimi, uzmanlaşmış tarafsız hakemlerin, sorunun yanlılıktan arınık olduğundan emin olmak için her soruyu detaylı bir biçimde incelendiği aşamadır. Yayın İncelemesi, Maddenin olabildiğince açık ve güvenilir olmasının denetlendiği safhadır. Bu safhada anlaşılırlık en önemli unsurdur. Paydaş İncelemesi aşaması ETS uzmanları tarafından yapılan tüm incelemelere ek olarak, üçüncü kişilerin, soruları teste dâhil olmadan önce gözden geçirmesinden oluşmaktadır. Pilot Uygulama basamağı yeni bir sorunun asıl hedef kitle üzerinde nasıl işlediğini görmek için yapılan deneme uygulamalarını kapsamaktadır. İstatistiksel değerlendirme adımında, maddelerin psikometrik özellikleri, istatiksel özellikleri, güçlü ve yanıtların öğrenciler arasında nasıl farklılaştığını görmek için küçük gruplara uygulanan test maddeleri analizlerinden oluşmaktadır. Kullanıma hazır hale getirme aşamasında, bir soru tüm standartları karşılamak ve nihayet ETS testinde bulunabilmek için her incelemeyi geçmek zorundadır. Bu maddeler nihai test formuna eklenmektedir. (https://www.ets.org/understanding_testing/test_development)

1.4. Amaç ve Önem

Bu bölümünde, tez çalışmasının amaçlarından bahsedilecek ve bu çalışmanın önemine değinilecektir.

1.4.1. Amaç

Öğrenci performansının hakemler tarafından değerlendirildiği durumları analiz etmek için geliştirilen HPM-SAK modeli, test geliştirme sürecindeki madde seçimi aşamasına uyarlanması durumunda kurulan hangi modelin daha uyumlu olduğu belirlenmeye çalışılmıştır.. Hiyerarşik puanlayıcı modelinde, hakem puanları, doğrudan öğrenci yeterliliğinin bir göstergesi olarak kabul edilmemektedir. Daha çok öğrenci performansının, ait olduğu örtük kategorinin bir göstergesi olarak tanımlanır.

Bu durum bir hiyerarşik yapıyı temsil eder. Örtük sınıf sinyal algılama modeli ise hakem hassasiyeti ve hakem sertliği-cömertliği gibi çeşitli etkilerin ölçülmesine imkân verir.

Bu araştırmada, öğrenci performansı yerine, yazarlar tarafından geliştirilen sorulara ait hakem puanları; öğrencinin ait olduğu örtük sınıflar yerine ise, maddelerin temsil ettiği özellik grupları incelenmiştir. Bu anlamda hiyerarşik örtük sınıf modeli, test geliştirme süreci için, soruları ölçtüğü özellikleri temsil etme düzeylerine göre kategorilendirebilecek ve aynı zamanda hakem sertlik ve cömertliğini sürece dâhil ederek test geliştirme sürecinde madde seçimine kullanılabilecek şekilde uyarlanmaya çalışılmıştır. Araştırmada modelin 3 ayrı düzeyinin hangisinin uyum iyiliğinin yüksek olduğu incelenmiştir.

1.4.2. Önem

Hakem kararlarına dayalı olarak öğrenci yeteneğini kestirmeye yönelik ölçümlerde hakem algısı üzerine yapılan çok sayıda araştırma literatürde yer almaktadır. Buna rağmen tamamen hakem kararları ve uzman görüşlerine dayandırılarak değerlendirilen ölçme süreçlerinden biri olan test maddesi niteliğini belirleme konusunda kullanılan yöntemler hakem güvenirliğini tek bir özellik olarak inceme eğilimindedir(Örn. Kappa, Genellenebilirlik Kuramı, Rash Modeli, Örtük Sınıf Yaklaşımı vb.). Bununla birlikte

puanlama anahtarı, kontrol listesi ya da rubrik gibi ölçüt anahtarları kullanıldığı durumlarda hakem algısının bir boyut olarak değerlendirildiği ve buna yönelik parametrelerin hesaplandığı çalışmalar dünya literatüründe de son zamanlarda yer almaya başlamıştır. Bu tez çalışmasının özgün niteliklerinden biri de ülkemizdeki bu tür çalışmaların ilklerinden biri olmasıdır.

Bununla birlikte araştırma HPM-SAK modele göre kurgulanan üç hiyerarşik düzeyin model uyumunu incelemiştir. Bu yönüyle de özgün bir değere sahiptir.

İncelenen modellerin kurgu biçimleri araştırmacılara pratik uygulamalarda madde seçimi ve test geliştirme süreci açısından kullanışlı ve uygulanabilir veriler sunmaktadır.

Pilot uygulama yapılmadan geliştirilen testlerde madde niteliğini belirlemek amacıyla öncelikli olarak hakem yargılarının kullanıldığı durumlarda (Örneğin, ÖSYM sınavları TEOG sınavları vb.) daha güvenilir madde puanları elde etme imkânı tanıyabileceği söylenebilir. Bu tür çalışmaların yaygınlaştırılması sınav kalitesinin arttırılmasına katkı sağlayacağı düşünülmektedir.

1.4.3. Problem Cümlesi

Madde niteliğinin belirlenmesi amacıyla kurulan HPM, iki düzeyli Sinyal Algılama Teorisi Hiyerarşik Puanlayıcı Modeli (HPM-SAK) ve üç düzeyli Sinyal Algılama Teorisi Hiyerarşik Puanlayıcı Modellilerinden (HPM-SAK) hangisinin uyum düzeyi daha yüksektir?

1.4.4. Alt Problemler

1. Test geliştirme sürecinde hakem yargılarına göre madde seçimi aşamasında kurulan HPM modelin uyum düzeyi nedir?
2. Test geliştirme sürecinde hakem yargılarına göre madde seçimi aşamasında kurulan 2 düzeyli HPM-SAK uyum düzeyi nedir?
3. Test geliştirme sürecinde hakem yargılarına göre madde seçimi aşamasında kurulan 3 düzeyli HPM-SAK uyum düzeyi nedir?

1.4.5. Sınırlılıklar

Bu araştırma 115K531 No'lu TÜBİTAK projesi kapsamında yürütülmüştür. Bu anlamda araştırma örneklemi ve veri topla araçları proje içeriği ile sınırlıdır. Geliştirilen soruların sadece matematik 6. Sınıf düzeyi olması araştırmanın bir başka sınırlılığını oluşturmaktadır.

Araştırmada kullanılan ölçekte her bir madde bağımsız olarak kabul edilmiştir. Aynı zamanda analizler ölçeğin tek boyutlu olduğu varsayımı ile gerçekleştirilmiştir. Bu durum araştırmanın bir diğer sınırlılığı olarak görülebilir.

1.4.6. İlgili Araştırmalar

T. Lawrence DeCarlo (1998), "Sinyal Algılama Kuramı ve Genelleştirilmiş Doğrusal Modeller" isimli makalesinde, genelleştirilmiş doğrusal modelleri, sürekli ve kategorik tepki değişkenleri için regresyon benzeri modellerin genel bir sınıfı olarak ele almıştır. Sinyal algılama modellerini, genelleştirilmiş doğrusal modellerin bir alt sınıfı olarak formüle edilebilir ve sonucu farklı temel dağılımlara dayanan zengin bir sinyal algılama modelleri olarak tanımlamıştır. Uç değer modelinin, birimden farklı eğimli normal veya lojistik ROC eğrilerinin örnekleri şeklinde yaygın olarak verilen, bazı klasik veri setleri için alıcı işleme karakteristiği (ROC) eğrilerini ürettiğini göstermiştir. Modellerin, tepki bağımlılıklarını tanımlamak, rasgele katsayılar eklemek veya daha genel temel alınan olasılık dağılımlarına izin vermek gibi çeşitli yollarla da genişletilebildiği göstermiştir.

Lawrence T. DeCarlo, Young Koung Kim, Matthew S. Johnson (2011), "Yapılandırılmış Yanıt Maddeleri İçin Sinyal Algılama Hiyerarşik Puanlayıcı Modeli" isimli makalelerinde, hiyerarşik puanlayıcı modeli (HPM) ile hakemlerin soru yanıtlarını puanlarken oluşan verilerin hiyerarşik yapısını açıklamışlardır. Bu yaklaşımda, hakem puanlarını, öğrenci yeterliklerinin doğrudan göstergesi olarak değil, daha ziyade sınav kalitesinin göstergeleri olarak ele almışlardır. Sınava giren kişinin yanıtlarının (örtük kategorik) niteliğini sırasıyla, kişinin yeterliliğinin bir göstergesi olarak ele almış ve hiyerarşik bir yapıya sahip olduğunu göstermişlerdir. Burada, sinyal algılama teorisi (SAK) ile elde edilen Örtük sınıf modelinin, bir puanlayıcı modeli olan HPM'nin ilk seviyesi için doğal bir aday olduğunu da göstermişlerdir. Örtük sınıf SAK modeli, basit

hakem sertlik veya cömertliğinin ötesinde hakem algısı ve hassasiyeti gibi çeşitli hakem etkilerini ölçmeyi sağlamıştır. HPM-SAK modelini büyük ölçekli bir değerlendirmeden elde edilen verilere uygulamış ve hakem performansının çeşitli yönlerinin bir özetini vermişlerdir.

Akihito Kamata (2001), "Hiyerarşik Gelleştirilmiş Doğrusal Modeller ile Madde Analizi" başlıklı makalesinde, Hiyerarşik genelleştirilmiş doğrusal modeli (HGLM) çok düzeyli bir madde tepki modelinin iki düzeyli bir formülü olarak sunmuştur. Bu makalesinde, HGLM'nin Rasch modeli ile eşdeğer olduğunu ve HGLM'nin karakteristiği olan kişinin kabiliyetinin, parametreler yerine rastgele etkiler şeklinde ifade edilebileceği göstermiştir. İki aşamalı madde analizi modelini, kişilik karakteristik değişkenleri ile gizli bir regresyon modeli olarak sunmuştur. Ayrıca, iki seviyeli HGLM modelinin sınıflar ve okullardaki gibi öğrencilerin performanslarının gruplar arası değişiminin araştırılmasına ve interaktif etkinin incelenmesine olanak tanıyan üç seviyeli gizli bir regresyon modeline genişletilebileceği göstermiştir.

Birgit Benkhoff (1997), "HPM Modelin İşveren ve Çalışanlar Yararına Bir Testi" isimli çalışmasında HPM nin güçlü versiyonlarının destekçisi tarafından yapılan bazı varsayımları test etmiştir. Özel olarak, yazar burada, HPM karakteristiği uyarlaması kullanarak taahhüt ile kalite ve esneklik arasındaki önerilen bağı araştırmıştır ve iş memnuniyeti, kalma niyeti ve örgütsel performans aralarındaki ilişkileri incelemiştir. Bu değişkeni iş doyumu ve kalma niyeti ile yakından ilgili bulmuştur ve en önemlisi örgütsel performansa önemli katkıda bulunduğunu görmüştür.

Banks, William P. (1970), yılında "Sinyal Algılama Kuramı ve İnsan Hafızası". İsimli çalışmasında, hafıza süreçleri çalışmasında sinyal algılama teorisinin (SAK) 4 kategorideki uygulamasını eleştirel bir biçimde incelemiştir. Bunlar: (a) bellek gücünü ölçeklendirmek, (b) unutulmayı işaret eden verilerin kritik yorumlaması, (c) iz depolamanın biçimini belirlemek ve tamamen öğrenme ya da hiç öğrenmeme sorununu çözmek ve (d) SAK'nin uzantılarını, hafızaya dayalı ayrım yapabilirliği daha düşük özümleme analizlerinde ölçeklendirmek şeklindedir.

2. YÖNTEM

Araştırmanın bu bölümünde araştırmanın türünden bahsedilecek, kullanılan ölçme araçları ve çalışma grubu hakkında bilgi verilecektir. Verilerin elde edilmesi ve analiz aşaması ayrıntılandırılacaktır.

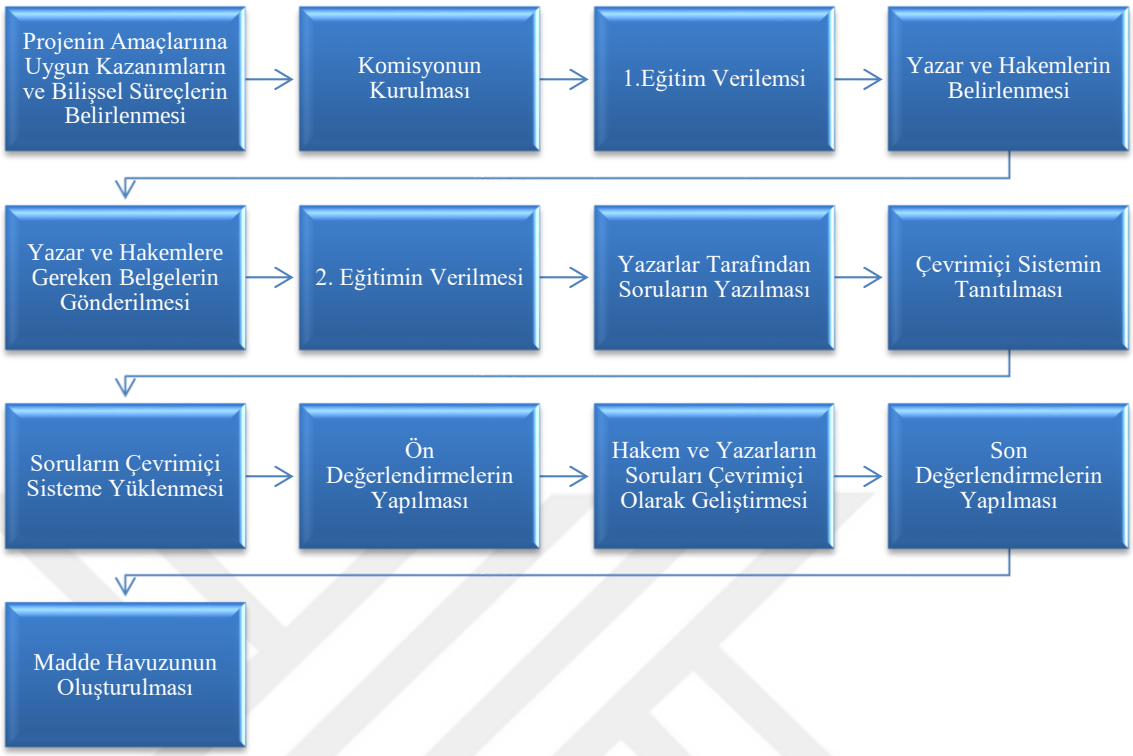
2.1. Araştırma Türü

Bu araştırma madde ve test geliştirme sürecinde hakem etkisini HPM-SAK modelle test etmek üzerine kurulduğundan temel bir araştırmadır. Temel araştırma, hipotez, teori veya yasaları formüle etmek ve test etmek amacıyla özellikleri, yapıları ve ilişkileri analiz eder. Temel araştırmalar, kuram (teori) geliştirmeyi ya da var olan kuramları sınamayı amaçlamaktadır (Karasar, 1998). Bu nedenle bu araştırma temel bir araştırma olarak değerlendirilebilir. Bunu yanında gerçek araştırma verilerinin model üzerinde test ediliyor olması nedeniyle de uygulamalı bir araştırma niteliği de taşımaktadır.

2.2. Verilerin Elde Edilmesi

Bu araştırmada 115K531 No'lu TÜBİTAK Projesi kapsamında elde edilen veriler kullanılmıştır. Bu verilerin elde edilmesi sürecindeki işlem basamakları, ölçme aracı, çalışma grubu, ölçme aracına ve hakemlere dair çeşitli betimsel istatistikler ve analiz aşaması bu bölümde ele alınmıştır.

Bu çalışma, yukarıda 115K531 No'lu TÜBİTAK Projesi kapsamında geliştirilen soruların psikometrik özelliklerini belirlemeye yönelik araştırmaları içermektedir. Projenin test geliştirme sürecinin iş-plan diyagramı aşağıda Şekil 5. ile verilmiştir.



Şekil 5. Projenin Süreç Diyagramı

Şekil 5. ile verilen diyagramda 115K531 No'lu TÜBİTAK Projesi kapsamında yürütülmüş süreçler verilmiştir. Aşamaların her biri hakkında genel bilgiler aşağıda verilmiştir. Bu tez çalışması bu süreçlerden yalnızca soruların çevrimiçi sisteme yüklenmesi aşaması ile yazarlar tarafından geliştirilen soruların, soru havuzuna eklenmesi aşaması arasındaki kısmı kapsamaktadır.

2.2.1. Soruların Çevrimiçi Sisteme Yüklenmesi

Projenin 2 aylık soru yazma süresi içerisinde soru yükleme süreci başlatılmıştır. Bu aşamada eğitim alan öğretmenler tarafından oluşturulan soru yazar ekibi ile projenin soru havuzunu oluşturma sürecinin ilk basamağı gerçekleştirilmiştir. Bu süreçte 91 öğretmen proje amaçları doğrultusunda yazmış olduğu soruları çevrimiçi sisteme yüklemeye başlamıştır.

Yazarların sadece soruyu yazmaları değil aynı zamanda proje amaçlarına göre sorunun hangi özellikleri ölçtüğü, hangi kazanımlar ile tanımlanabileceği ve bilişsel süreç becerileri ya da üst düzey düşünme becerileri ile nasıl ilişkilendirilebileceğine dair

bilgileri de tanımlamaları istenmiştir. Bu tanımlamalar proje kapsamında geliştirilen soruların daha sonra oluşturulacak Q matrisler ile ilişkilendirilmesinin ön çalışması olmuştur. Öğretmenler sürecin sonunda yazılmış olan 325 soru aşamalı olarak sisteme yüklenmiştir. Bu sayede sorular yüklendikçe ilgili hakemlere yönlendirilmiş ve hakemler için soru değerlendirme süreci de zaman anlamında rahatlatılmıştır. Soru türleri şu şekildedir;

Açık uçlu(1): Öğrenciden açıklama beklenen, kısmi cevaplanabilen, işlem basamaklarının görülmesi istenen türdeki sorular,

Çoktan seçmeli(2): 4 veya 5 seçenekli çoktan seçmeli sorular,

Kısa cevaplı(3): Cevabın sadece bir sayı ya da birkaç kelime istendiği durumlarda kullanılan sorular,

Doğru yanlış(4): Birden fazla doğru veya yanlış ifade ile oluşturulan sorular.

115K531 No'lu TÜBİTAK Projesi kapsamında yazılan soruların hangi türlerde ve kaçar tane olduğu Tablo 2. de verilmiştir.

Tablo 2. Yazılan Soru Türleri ve Soru Türlerine Ait Soru Sayısı

Soru türü	Soru türü kodu	Soru sayısı
Açık uçlu	1	119
Çoktan seçmeli	2	62
Kısa Cevaplı	3	114
Doğru yanlış	4	30
Toplam		325

Tablo 2. de görüldüğü gibi ağırlıklı olarak açık uçlu ve kısa cevaplı sorular yazılmıştır. 119 açık uçlu, 114 kısa cevaplı soru sisteme yüklenmiştir. Bunu Çoktan seçmeli ve Doğru yanlış türündeki sorular takip etmiştir. 62 çoktan seçmeli, 30 doğru yanlış türünde soru sisteme yüklenmiştir.

2.2.2. Proje Sorularını Değerlendirme Süreci

Hakem–yazar arasındaki anlık bilgi ve doküman alışverişi süreci geleneksel soru değerlendirme yaklaşımından soruyu geliştirme anlayışına çevirmiştir. Yazar ve hakem

arasındaki bilgi paylaşım süreci sadece iki kişi arasında olmakla birlikte kişisel bilgiler yönünden iki taraf için de tamamen anonim olmuştur. Her hakem soruyu ön ve son değerlendirme olarak iki kere puanlamıştır.

Ön değerlendirme hakemin soru yazarı ile görüşmesinden önce gerçekleşmekteyken son değerlendirme yazar ile hakem arasındaki görüşmeler sona erdiğinde hakem tarafından yapılmıştır.

Hakemler soru değerlendirmelerini, sistem üzerindeki puanlama anahtarı aracılığı ile proje kuruluna göndermişlerdir. Bu sayede sorular hem amaca uygun olarak hızlı ve kör hakemlik kurallarıyla değerlendirilmiş hem de sorulara ilişkin değerlendirme verileri dijital olarak toplanabilmiştir.

2.2.3. Proje Sorularının Ön Değerlendirmelerinin Yapılması

Bu süreçte değerlendirme aşamasına ilişkin ilk çalışmalar başlatılmıştır. Projeye ait web sitesi (www.cdmproject.org) projenin tanıtımının yanı sıra geliştirilen soruların değerlendirme sürecinin organizasyonuna da hizmet etmektedir.

Hakemlerin kendi sistemlerine yüklenen soruları inceleyip, yazarlarla görüşmelerinin gerçekleşmesinden önce, hazırlanmış olan puanlama anahtarı yardımıyla puanlar vererek sorunun ilk haline ilişkin ön değerlendirmeleri yapma aşaması başlatılmıştır.

Burada hakemler kime ait soruyu değerlendirdiklerini, değerlendirmenin hiçbir aşamasında bilmedikleri için anonimlik korunmuştur. Her hakem ortalama 35 soru değerlendirmiştir. Sorunun ilk haline ilişkin ön değerlendirmeler tamamlandıktan sonra soruların yazarlar ile aktif bilgi paylaşımında bulundukları, soru geliştirilme sürecine geçilmiştir. Puanlama anahtarı ve ilgili açıklamalar "Ölçme Aracı" başlığında ayrıntılandırılacaktır.

2.2.4. Hakem ve Yazarların Soruları Çevrimiçi Olarak Geliştirmesi

Hakemler ön değerlendirmeleri tamamladıktan sonra, çevrimiçi sistemde yazarlara görüşlerini bildirmiş, gerekliği olduğu noktada çeşitli dokümanlar paylaşmış ve sorunun geliştirilmesi için önerilerde bulunmuştur. Soru yazarları sorularını sisteme yükledikten

sonra süreçten kopmamış ve soru değerlendirme sürecine de aktif olarak katılarak yazdıkları soruların daha üst düzey nitelikleri ölçebilmesi amacıyla geliştirilmesine katkı sağlamışlardır. Hakem ve yazarın karşılıklı yazışmalarını sürdürerek soru üzerinde fikir alış verişi yapabildikleri çevrimiçi sistem sayesinde sorular daha hızlı şekilde son haline getirilebilmiştir. Soruların düzenlenmesi ve geliştirilmesi tamamlanana dek bu karşılıklı fikir alış verişi devam ettirilmiş bu süreçte hakem ve yazar işbirliği ile soru kalitesi arttırılarak sorunun proje amaçlarına hizmet edebilecek son haline getirilmesi sağlanmıştır.

2.2.5. Proje Sorularının Son Değerlendirmelerinin Yapılması

91 matematik öğretmeni tarafından 6. sınıf düzeyinde yazılan toplam 325 proje sorusu 28 hakem tarafından proje web sitesinde çevrimiçi değerlendirilmiştir. Bu süreçte her soru 3 ayrı hakem tarafından değerlendirildiği için toplam 975 ön ve son değerlendirme yapılmış, toplam 28 hakem ortalama 35 ayrı soruyu değerlendirmiştir. Soru değerlendirme süreci çevrimiçi gerçekleştirilmiştir. Son değerlendirmeye ait, her bir hakeme anonimlik amacı ile atanan hakem numaraları ve hakemlerin değerlendirdiği soru numaraları EK1.'de, her hakemin değerlendirdiği soru sayıları ve her yazarın yazdığı soru sayıları EK2.'de listelenmiştir. Soru yazarı ve hakem arasında anonimlik değerlendirmenin her aşamasında korunmuştur.

Soruların hakem ve yazar işbirliği ile soru kalitesi arttırılarak sorunun proje amaçlarına hizmet edebilecek son haline getirilmesinin ardından, son değerlendirme aşamasına gelinmiştir. Bu aşamada hakemler sorunun son halini göz önünde bulundurarak ön değerlendirme aşamasında kullanılan puanlama anahtarı yardımıyla sorunun son haline de puanlar vermiştir.

Projedeki soru kalitesini ölçen puanlama anahtarı geliştirilmiştir. Geliştirilen ölçek Tablo 3.'te verilmiştir.

Tablo 3. Puanlama Anahtarı

	Değerlendirme Ölçütleri	**Hiç**	**Yetersiz**	**Kullanılabilir**	**Uygun**	**Tamamen uygun**
1	Soru kazanımın düzeyine uygundur.					
2	Sorunun çözümü birden çok matematiksel yeterlilik/davranış/beceri gerektirmektedir.					
3	Sorunun anlatım dili 6. sınıf öğrencileri için açık ve anlaşılırdır.					
4	Soru üst düzey düşünme becerilerini ölçebilecek niteliktedir.					
5	Soru kültürel, coğrafi ve etnik yanlılık içermiyor.					
6	Tercih edilen soru formatı (kısa cevaplı- açık uçlu vs) uygundur.					
7	Soru sadece ezber bilgisi kullanarak çözülebilir.					
8	Soru PISA ve TIMSS sorularına benzer niteliktedir.					
9	Soru proje kapsamında ölçülmek istenen ilgili kazanımları ölçebilecek niteliktedir.					
10	Soru bağlamı gerçek yaşam problemlerine yönelik hazırlanmıştır.					
11	Soruda kullanılan görsel materyaller soru için gereklidir.					
12	Soruda kullanılan senaryo/tablo/resim/şekil vb. uygundur.					
13	Soru bağlamı matematiksel becerileri kullanmaya uygundur.					
14	Soru açık, belirgin ve cevaplanabilir bir sorudur.					
15	Soru özgündür.					

*****Yukarıdaki puanlama anahtarı 115K531 No'lu TÜBİTAK Projesi kapsamında geliştirilmiştir.

Tablo 3.'te verilen maddeler yazar tarafından geliştirilen sorulara ilişkin hakem değerlendirmesi sürecinde kullanılan puanlama anahtarını oluşturmaktadır. Hakemler değerlendirdikleri soruları çevrimiçi sistem içinde yukarıdaki maddeleri 0 ile 4 arasında puanlayarak gerçekleştirmişlerdir.

2.2.6. Soru Havuzunun Oluşturulması

Yukarıda anlatılan değerlendirme süreci sonrasında hakemlerden en yüksek puanı alan soruların projede kullanılmak üzere soru havuzuna eklenmiştir. Ön test ve izleme testleri

süresince uzmanlar tarafından kapsam ve test içeriği yönünden soru havuzunda bulunan 325 soru 280'e indirilmiştir.

115K531 No'lu TÜBİTAK Projesine ait aşamaların, soru havuzuna eklenme aşamasına kadar olan kısmı yukarına anlatılmıştır. Soruların deneme uygulamalarına hazırlanması şeklinde devam eden süreç bu çalışmada yer almamaktadır. Yukarıda da bahsedildiği gibi bu tez çalışması süreçlerden yalnızca soruların çevrimiçi sisteme yüklenmesi aşaması ile soruların, soru havuzuna eklenmesi aşaması arasındaki kısmı kapsamaktadır.

2.3. Ölçme Aracı

Bu tez çalışmasında kullanılan ölçme aracı proje kapsamında hakemlerin soru kalitesini belirlemek için kullandığı 15 maddeden oluşan ve 0 ile 4 arasında puanlanan beşli bir puanlama anahtarıdır. Bu puanlama anahtarı ile proje yazarlarının yazdığı sorular, hakemler tarafından değerlendirilmiştir. Bir sorunun bir değerlendirmeden alabileceği madde skoru 0 ile 60 arasında değişmektedir. Puanlama anahtarı ile hakemler tarafından sorulara verilecek puanlar için dikkat edileceklerle ilgili açıklama aşağıdaki Tablo 4.'te listelenmiştir.

Tablo 4. Puanlama Anahtarı İçindeki Değerlendirme Ölçütleri ve Açıklamaları

	Değerlendirme ölçütleri	**Açıklama**
1	Soruda tanımlanan kazanımlar soruyla uyumludur.	Soru yazarının soru için tanımladığı kazanımlar sorunun içeriği ile tutarlı mı? Belirlenen kazanımlar incelendiğinde eksik, yanlış ya da fazla kazanım bulunmakta mı?
2	Sorunun çözümü birden çok matematiksel yeterlilik/davranış/beceri gerektirmektedir.	Soruyu çözen öğrenci doğru cevaba ancak birden fazla matematiksel beceri ve/veya kazanımı kullanarak mı ulaşıyor? Öğrencinin soruyu çözebilmesi için soruda tanımlanan kazanımların hepsine sahip olması gerekli mi?
3	Sorunun anlatım dili 6. sınıf öğrencileri için açık ve anlaşılırdır.	Soruda kullanılan dil ve ifadeler 6. Sınıf öğrencilerinin rahatça anlayabileceği düzeyde mi? Soru dil ve anlatım bakımından imla ve yazım kurallarına uygun mu?
4	Soru üst düzey düşünme becerilerini ölçebilecek niteliktedir.	Soru; karar verme, yeni bir düşünce, yordama, rutin olmayan problemleri çözme, eleştirel düşünme, kanıta dayalı düşünme, yaratıcı düşünme gibi üst düzey davranışları yokluyor mu?

		Soru, öğrencinin kendi çözümünü kontrol etmeyi ve kendi çözüm sürecini düzenlemeyi gerektiriyor mu? Soru, verilen durumu analiz ederek buna uygun bir çözüm stratejisi bulmayı gerektiriyor mu? Sorunun çözümüne ulaşmak yoğun bir zihinsel çaba ve muhakeme gerektiriyor mu? Soru, yeni strateji ve yaklaşım oluşturmayı gerektiriyor mu?
5	Soru kültürel, coğrafi ve etnik yanlılık içermiyor.	Soru 6. Sınıf öğrencilerin ait oldukları alt kültür, etnik kökenleri, cinsiyetleri veya yaşadıkları bölgeyle ilgili avantaj ya da dezavantaj içeriyor mu? Soruda ölçülen özellikler bakımından aynı düzeyde olan iki öğrenciden birisi yukarıda söz edilen gruplardan birine dahil olduğu için soruyu diğerine göre daha hızlı çözebilir mi ya da bu durum onun doğru cevaplama olasılığını arttırır mı?
6	Tercih edilen soru formatı (kısa cevaplı- açık uçlu vs.) uygundur.	Soru yazarı soruda ölçülen özellikler bakımından en uygun formatı tercih etmiş mi? Soru başka bir formatta ilgili kazanımları daha iyi ölçebilir mi? Soru formatı değiştirilerek sorunun ölçtüğü kazanımın düzeyi arttırılabilir mi?
7	Soru sadece ezber bilgisi kullanarak çözülebilir.	Soru; sadece daha önce öğrenilen kural, formül, tanım ve benzeri bilgileri hatırlamayı veya ezberlemeyi mi ölçer? Öğrenci soruyu daha önceden çözdüğü sorulara benzerliğini kullanarak ezber bilgisiyle çözebilir mi? Soru; Kural, formül, tanım vb. bilgilerin kavramsal anlamına yönelik bir bağlantı kurmayı gerektiriyor mu?
8	Soru PISA ve TIMSS sorularına benzer niteliktedir.	Soru biçim ve içerik olarak PISA ve TIMSS sınavları sorularına ne kadar benzemektedir? Soru PISA ve TIMSS soruları yerine kullanılabilir mi?
9	Soru proje kapsamında ölçülmek istenen ilgili kazanımları ölçebilecek niteliktedir.	Soru 6. Sınıf öğrencileri için uygun kazanımları içermekte mi? Üst düzey kazanımları ölçmeye uygun mu? Bu gibi sorular ile karşılaşan öğrenci PISA ve TIMSS sınavlarında diğer öğrencilerden daha başarılı olabilir mi?
10	Soru bağlamı gerçek yaşam problemlerine yönelik hazırlanmıştır.	Soru ile yoklanan problem durumu 6. Sınıf düzeyi öğrenciler için gerçek yaşamda karşılaşabilecekleri bir içeriğe sahip mi? Soru öğrencilerin anlayabilecekleri veya ilgilerini çekebilecek bir bağlama sahip mi?
11	Soruda kullanılan görsel materyaller soru için gereklidir.	Soruda yer alan görsel materyal sorunun çözümü için gerekli midir? Kullanılan görsel soruyu daha anlaşılır ya da ilgi çekici yapmakta mıdır?

12	Soruda kullanılan senaryo/tablo/resim/şekil vb. uygundur.	Sorudaki görsel materyal veya senaryo amaca uygun kullanılmış mı? Sorunun biçimsel açıdan geliştirilmeye ihtiyacı var mı? Kullanılan görsel materyal veya senaryo sınıf düzeyine uygun mu?
13	Soru bağlamı matematiksel becerileri kullanmaya uygundur.	Sorunun kökü ve kurulan bağlam ölçülen matematiksel beceriler ile ilişkili mi? Öğrenci soru bağlamını anlamadan sadece rakamları ve şekilleri kullanarak doğru cevabı bulabilir mi? Soru bağlamı gerçek dünyadaki problemlerin altında yatan matematiksel değişkenleri ve yapıları tanımlama ve kullanılabilecek varsayımları oluşturmaya uygun mu? Soru bağlamı gerçek hayattaki bilgilerin matematiksel bir gösterimini oluşturmaya uygun mu? Soru bir gerçek yaşam problemini sembolik dil kullanarak gösterebilmek için uygun değişkenleri, sembolleri, diyagramları ve standart modelleri kullanmaya uygun mu?
14	Soru açık, belirgin ve cevaplanabilir bir sorudur.	Sorunun çözümü için gerekli bilgiler kökte eksiksiz verilmiş mi? Öğrencinin kökteki bilgilerden yararlanarak çıkarım yapması gerekiyorsa yapılacak çıkarımlar sınırlandırılmış mı? Öğrenci farklı bir açıdan soruya yaklaşarak farklı ve beklenmeyen sonuçlara ulaşabilir mi?
15	Soru özgündür.	Soru biçim, içerik, bağlam, çözüm stratejileri açısından özgünlük taşıyor mu? Soru çok bilinen bir yöntemle ya da alışıla gelmiş bir çözüm yoluyla çözülebilir mi? Soru daha önce kullanılan bir soruyla içerik ve çözüm yolu açısından ciddi benzerlik taşıyor mu?

*****Yukarıdaki puanlama anahtarı 115K531 No'lu TÜBİTAK Projesi kapsamında geliştirilmiştir.

Tablo 4. Puanlama anahtarının kullanımını kolaylaştırmaya yönelik ifadeler ve açıklamaları içermektedir. Her hakem bir soruyu değerlendirirken yukarıda verilen puanlama anahtarını ve her maddeye ilişkin açıklamaları kullanmaktadır.

Puanlama anahtarındaki her madde, değerlendirilen sorunun açıklamalarda verilen ifadeler açısından yeterlik düzeylerine göre 0 ile 4 arasında 5 çeşit puan alabilmektedir. Yazarlar tarafından yazılmış olan her sorunun, hakem değerlendirmesi sonunda bir sorunun toplam skoru 0 ile 60 arasında değişmektedir.

Soru niteliklerini belirlemek için geliştirilen 15 maddelik puanlama anahtarı kullanılarak yapılan 688 değerlendirmeye ait güvenirliği (Cronbach Alfa) ve Standart Cronbach Alfa hesaplanmış ve aşağıdaki Tablo 5.'te verilmiştir.

Tablo 5. Güvenirlik İstatistikleri

Cronbach Alfa	Standart Cronbach Alfa	Madde Sayısı
0,907	0,917	15

Tablo 5.'te görüldüğü gibi 15 maddeden oluşan puanlama anahtarına ait alfa değeri 0,907 ve standart alfa değeri 0,917 bulunmuştur. Bu ölçme aracının oldukça güvenilir olduğuna işaret etmektedir.

2.4. Çalışma Grubu

Araştırmanın çalışma grubu 115K531 No'lu TÜBİTAK Projesi tarafından geliştirilen 325 soru ve 28 Hakem içinden seçilmiştir. Projede yer alan hakemlerin bazıları hiç soru değerlendirmemiş veya yalnız bir soru değerlendirmiş, kimi sorular da ön değerlendirme aşamasından geçememiş olduğu için bu hakemler ve sorular bu çalışmanın örnekleminde dâhil edilmemiştir. Bu çalışanın örnekleminde 280 soru ve 23 hakem ile oluşturulmuştur. Toplam 688 değerlendirme kullanılmıştır.

2.4.1. Sorulara İlişkin Betimsel İstatistikler

Çalışmanın bu kısmında yukarıda bahsedilen, proje kapsamında geliştirilmiş soru kalitesini belirlemede kullanılan puanlama anahtarında bulunan maddelerin ve bu puanlama anahtarı yardımıyla hakemlerin değerlendirdikleri soruların betimsel istatistiklerine yer verilmiştir.

Araştırma örnekleminde kullanılan sorular yukarıda da daha önce belirtildiği gibi 119 açık uçlu, 62 çoktan seçmeli, 114 kısa cevaplı ve 30 doğru yanlış tipindeki toplam 325 soru içinden seçilen 280 soru ile oluşturulmuştur. Her soru oluşturulan 15 soruluk puanlama ölçeği ile 0 ila 60 arasında puan alabilmektedir.

Yazarlar tarafından geliştirilen sorulara verilen kodlar ve bu soruların 3 ayrı hakemden, geliştirilen puanlama ölçeğinden aldığı puan ortalamamaları Tablo 6.'da verilmiştir.

Tablo 6. Soruların Puan Ortalamaları

Soru kodu	**Ortalama**	**Soru kodu**	**Ortalama**	**Soru kodu**	**Ortalama**	**Soru kodu**	**Ortalama**	**Soru kodu**	**Ortalama**	**Soru kodu**	**Ortalama**	**Soru kodu**	**Ortalama**
1010	52,00	2234	49,67	3729	29,50	5064	54,67	6239	37,50	7339	45,00	8558	38,50
1113	53,00	2252	51,33	3737	38,67	5073	44,00	6251	52,50	7372	50,00	8599	52,50
1119	59,00	2363	55,00	3753	38,75	5077	44,67	6259	36,50	7548	46,33	8642	38,00
1151	47,50	2384	42,00	3781	46,67	5106	48,33	6323	30,50	7570	49,75	8675	44,00
1259	46,00	2391	32,00	3846	33,50	5128	47,00	6336	52,33	7586	32,67	8724	51,50
1267	48,67	2481	34,50	3873	52,00	5158	56,00	6360	39,33	7624	54,00	8867	42,00
1289	48,50	2526	47,00	3908	34,50	5199	53,33	6441	40,33	7650	48,00	8875	34,50
1320	48,00	2557	43,67	3915	31,00	5300	39,33	6462	52,00	7682	54,67	8895	50,33
1332	39,50	2567	51,67	3928	50,00	5321	50,00	6466	47,00	7705	46,50	8916	32,50
1346	36,67	2628	41,33	3940	45,33	5403	53,67	6470	48,50	7725	43,67	8943	35,50
1356	55,50	2668	46,00	3960	35,50	5462	54,50	6471	47,00	7729	52,00	8948	41,00
1388	50,50	2805	48,33	3999	46,50	5472	41,50	6493	48,50	7785	35,00	8951	26,33
1400	47,67	2828	53,00	4024	57,50	5486	35,00	6526	41,00	7823	55,33	8992	41,67
1536	47,00	2976	36,50	4045	45,00	5633	27,33	6564	40,33	7864	51,00	9049	33,50
1550	47,67	2985	21,50	4163	47,00	5637	21,67	6584	45,00	7868	37,00	9065	52,00
1553	43,00	3013	34,67	4370	48,33	5648	21,50	6659	43,50	7875	49,00	9074	52,00
1616	50,00	3019	28,00	4420	51,33	5678	48,50	6725	48,67	7914	49,00	9077	49,33
1621	33,00	3047	44,00	4427	50,00	5697	42,00	6728	55,00	7925	47,00	9101	34,67
1626	45,50	3136	53,50	4442	44,00	5720	43,33	6737	51,00	7943	35,33	9123	33,00
1634	51,00	3138	52,00	4555	51,50	5728	44,67	6746	49,67	7948	55,00	9157	44,00
1650	43,50	3159	47,67	4590	50,00	5737	32,50	6757	41,50	7965	46,00	9270	41,00
1664	51,00	3198	49,50	4650	22,50	5745	30,33	6758	49,50	8019	49,67	9297	36,00
1788	54,50	3281	43,33	4669	42,00	5752	50,00	6766	45,50	8077	49,33	9353	41,50
1797	56,50	3283	56,00	4718	46,67	5756	37,00	6830	38,00	8147	47,50	9408	36,50
1834	53,50	3328	44,50	4732	53,67	5770	34,33	6858	23,50	8167	37,00	9430	50,00
1863	49,00	3352	42,50	4756	41,33	5821	33,33	6876	47,67	8198	47,67	9442	30,00
1886	53,50	3370	40,50	4794	47,50	5827	42,00	6902	47,00	8228	48,00	9464	42,33
1901	20,50	3380	52,00	4831	42,00	5838	48,00	6912	43,67	8248	28,00	9507	53,50
1979	52,33	3381	54,00	4838	45,67	5856	31,67	6944	30,50	8250	21,50	9579	57,00
1988	40,00	3383	47,67	4848	34,67	5898	56,00	6963	41,50	8287	47,00	9610	55,00
1998	56,00	3384	53,00	4945	49,50	5918	45,00	7012	35,33	8288	32,00	9624	53,67
2016	35,67	3385	48,00	4950	44,33	5936	36,00	7017	51,67	8318	52,50	9739	55,00

2018	42,33	3386	54,50	4971	29,20	5953	39,00	7023	52,00	8342	51,00	9777	39,33
2095	47,00	3388	56,00	4978	45,33	6050	53,50	7026	40,00	8349	48,33	9796	44,00
2109	24,00	3438	34,00	4992	48,00	6056	44,50	7120	45,50	8366	43,00	9803	44,50
2114	32,50	3445	48,50	5017	37,00	6082	51,50	7197	42,33	8435	50,00	9830	53,00
2116	49,50	3455	49,50	5036	41,33	6096	48,00	7210	48,00	8466	52,00	9891	53,50
2117	41,33	3456	31,50	5045	38,33	6145	55,00	7215	43,50	8492	41,50	9911	36,00
2147	52,67	3492	41,67	5050	48,50	6217	42,67	7280	56,00	8549	39,67	9914	43,00
2206	50,00	3690	57,00	5057	53,50	6221	32,50	7305	54,67	8554	37,50	9919	49,67

Tablo 6.'da görüldüğü gibi 280 soru 20,5 ile 59 arasında bir ortalama soru skoruna sahiptir. Soru skoru [20, 30) aralığında olan 14 soru, [30, 40) aralığında olan 59 soru, [40, 50) aralığında olan 125 soru ve [50, 60) aralığında olan 82 soru bulunmaktadır.

Tablo 7.'de, hakemlerin kullandığı puanlama anahtarının maddeler için ortalamaları, standart sapmaları, madde test korelasyonları ve madde çıktığı zaman Cronbach Alfa güvenirlik katsayısındaki değişim verilmiştir.

Tablo 7. Madde İstatistikleri

Soru numarası	Soru içeriği	Ortalama	Standart Sapma	Madde-Test Korelasyonu	Madde silindiğinde elde edilecek Cronbach Alfa
1	Soru kazanımın düzeyine uygundur.	3,190	1,032	0,594	0,901
2	Sorunun çözümü birden çok matematiksel yeterlilik/davranış/beceri gerektirmektedir.	2,968	1,091	0,726	0,896
3	Sorunun anlatım dili 6. sınıf öğrencileri için açık ve anlaşılırdır.	3,105	1,041	0,684	0,898
4	Soru üst düzey düşünme becerilerini ölçebilecek niteliktedir.	2,824	1,069	0,790	0,894
5	Soru kültürel, coğrafi ve etnik yanlılık içermiyor.	2,413	1,641	0,144	0,924
6	Tercih edilen soru formatı (kısa cevaplı- açık uçlu vs.) uygundur.	3,128	1,070	0,647	0,899
7	Soru sadece ezber bilgisi kullanarak çözülebilir.	3,012	1,198	0,038	0,920

8	Soru PISA ve TIMSS sorularına benzer niteliktedir.	2,762	1,170	0,799	0,893
9	Soru proje kapsamında ölçülmek istenen ilgili kazanımları ölçebilecek niteliktedir.	2,903	1,152	0,798	0,893
10	Soru bağlamı gerçek yaşam problemlerine yönelik hazırlanmıştır.	3,015	1,120	0,734	0,896
11	Soruda kullanılan görsel materyaller soru için gereklidir.	3,029	1,242	0,554	0,902
12	Soruda kullanılan senaryo/tablo/resim/şekil vb. uygundur.	3,010	1,172	0,680	0,898
13	Soru bağlamı matematiksel becerileri kullanmaya uygundur.	2,985	1,039	0,821	0,893
14	Soru açık, belirgin ve cevaplanabilir bir sorudur.	3,045	1,130	0,633	0,899
15	Soru özgündür.	2,975	1,128	0,692	0,897

*****Yukarıdaki puanlama anahtarı 115K531 No'lu TÜBİTAK Projesi kapsamında geliştirilmiştir.

Tablo 7.'de görüldüğü gibi puanlama anahtarının maddeler için ortalamaları 2,413 ile 3,19 arasında değişmektedir. Madde standart sapmaları 1,032 ile 1,641 arasında değişmektedir. En yüksek madde test korelasyonu 13.madde ile test arasında ve değeri 0,821 iken en düşük madde test korelasyonu 7. Madde ile test arasında ve değeri 0,038 olarak hesaplanmıştır.

Maddeler çıkartıldığı zaman elde edilecek Cronbach Alfa değeri 0,893 ile 0,924 arasında değişmektedir. Tüm maddelerin bulunduğu puanlama anahtarının Cronbach Alfa değeri 0,907 olduğundan 7. ve 5. Maddelerin puanlama anahtarından çıkarılmasının güvenirliği sırasıyla 0,013 ve 0,017 değerinde arttıracağı görülmektedir. Diğer maddelerin çıkarılması güvenirlik değerini az da olsa düşürmektedir. Güvenirlik değerindeki en büyük azalışın 9. Maddenin çıkarılması ile ve ilk değere göre 0,014 değerinde bir azalma oluşacağı görülmektedir. Ayrıca tüm maddeler için Test istatistikleri de hesaplanmıştır. Puanlama ölçeği yardımıyla her soruya verilmiş olan puanların toplamı ile oluşturulan

madde skorlarının ortalaması ve bu ortalamaya ait standart hata, en düşük puanların %5'inin kesilmesi ile oluşan ortalama, medyan, standart sapma, minimum-maksimum puan, açıklık ve çeyrekler açıklığı gibi istatistikler hesaplanmış, Tablo 8.'de listelenmiştir.

Tablo 8. Test İstatistikleri

İstatistik	Değer	Standart Hata
Ortalama	44,363	0,438
5% Kesilen Ortalama	45,226	
Medyan	47,000	
Varyans	131,766	
Standart sapma	11,479	
Minimum puan	4,000	
Maximum puan	60,000	
Açıklık	56,000	
Çeyrekler Açıklığı	14,000	

Tablo 8.'de görüldüğü gibi puanlama anahtarı ile değerlendiren soruların aldıkları madde skorlarının ortalamaları 44,363'tür. En düşük puanları alan soruların %5'i çıkarıldığında oluşan ortalama 45,226'dır. Ortalamada ciddi bir değişimin olmaması en düşük puan alan soruların da ortalamaya yakın puanlar aldığına işaret etmektedir.

Madde skorlarına ait medyan 47,000, varyans 131,766 ve standart sapma 11,479'dır. Minimum madde puanı 4,000 iken maksimum madde puanı 60,00'dır ve bu durumda açıklık 56,000 değerini alır ayrıca çeyrekler açıklığı da 14,000'tür.

2.4.2. Hakemlere İlişkin Betimsel İstatistikler

115K531 No'lu Proje kapsamında çalışan 28 hakemin, örneklem için uygun puanlama yapanlarından 23 tanesi bu çalışmada kullanılmıştır. Hakemler toplam 688 değerlendirme yapmışlardır. Aşağıdaki Tablo 9.'da her bir hakemin kaç tane soru değerlendirdiği, değerlendirdiği sorulara verdiği puanların ortalaması ve standart sapması verilmiştir.

Tablo 9. Hakemlere ait Soru Sayısı, Puan Ortalaması ve Standart Sapma

		Hakemin yaptığı değerlendirme sayısı	Ortalaması	Standart sapması
Hakemler	Hakem 11	25	39,48	7,16
	Hakem 12	28	31,50	8,74
	Hakem 13	29	46,21	8,00
	Hakem 15	30	33,50	18,12
	Hakem 16	24	51,13	5,16
	Hakem 17	26	42,92	15,35
	Hakem 18	28	46,36	7,89
	Hakem 19	10	33,30	20,23
	Hakem 20	33	40,67	11,18
	Hakem 21	34	38,15	10,28
	Hakem 22	32	44,09	5,85
	Hakem 23	35	37,14	10,60
	Hakem 24	30	49,47	3,58
	Hakem 25	34	45,94	7,91
	Hakem 26	31	54,42	4,44
	Hakem 27	30	31,83	10,63
	Hakem 29	31	53,00	7,77
	Hakem 31	34	49,62	6,09
	Hakem 32	36	48,86	9,74
	Hakem 33	31	56,19	4,87
	Hakem 35	34	40,03	3,57
	Hakem 36	31	51,55	7,58
	Hakem 38	32	47,06	7,43
	Toplam	688		
	Ortalama	29,913		

Tablo 9.'da görüldüğü gibi hakemler en az 10 en çok 36 değerlendirme yapmıştır. Her hakemin yaptığı ortalama değerlendirme sayısı yaklaşık 30'dur. Hakemlerin verdiği ortalama puanlar 31,5 ile 56,19 arasında değişmektedir. Bu puan ortalamalarına bakılarak en katı hakemin Hakem 12, en cömert hakemin de Hakem 33 olduğu söylenebilir. Bu puanlara dair standart sapmalar da 3,57 ile 20,23 arasındadır.

2.5. Bulguların Elde Edilmesi ve Analiz

Araştırmada HPM-SAK model için R yazılımı ve R kodları kullanılmıştır. Özellikle Alexander Robitzsch tarafından geliştirilmiş R kodları kullanılmıştır. EK3.'de kodlara ait ayrıntılı bilgi verilmiştir. Bu çalışmada kullanılan algoritma aşağıda verilen istatistikle hesaplama yapmaktadır.

Modelin özellikleri DeCarlo ve diğerleri (2011) makalesindeki gibidir. İki düzeyli ideal puanlayıcı modeli, p kişisi, i maddesi, $\eta = 0, \dots, K$ için,

$$P(\eta_{pi} = \eta | \theta_p) \propto \exp(a_i q_{ik} \theta_p - \tau_{ik})$$

Şeklindedir. Birinci düzey, p kişisinin i maddesi ve r hakemi için X_{pir} puanı Sinyal algılama modeli ile şu şekilde modellenir.

$$P(X_{pir} \leq k | \eta_{pi}) = G(c_{irk} - d_{ir}\eta_{pi})$$

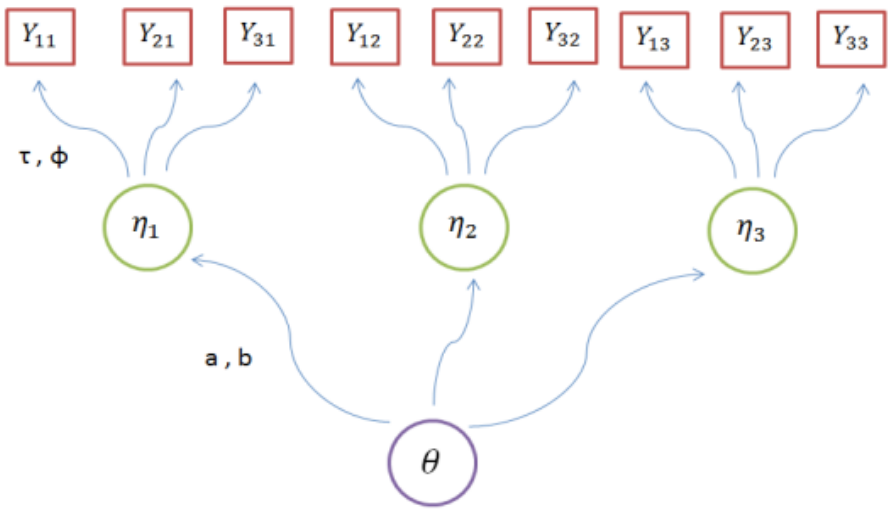
Burada G lojistik dağılım fonksiyonu ve kategoriler $k = 1, \dots, K$ dır. Ayrıca madde tepki kuramı buna denk olarak şu şekilde de yazılabilir.

$$P(X_{pir} \geq k | \eta_{pi}) = G(d_{ir}\eta_{pi} - c_{irk})$$

c_{irk} eşitlikleri ayrıca $c_{irk} = c_k$ (est.c.rater= 'e'), $c_{irk} = c_{ik}$ (est.c.rater= 'i') ve $c_{irk} = c_{ir}$ (est.c.rater= 'r') ile sınırlanabilir. Aynı durum hassasiyet parametreleri için de geçerlidir.

Bu R kodları kullanılarak yapılan analizler ile araştırmada geliştirilen puanlama anahtarı yardımıyla yapılan değerlendirmelerde madde seçim aşamasında Hakemlerin yanlılıklarını, hassasiyetini, sertlik ve cömertlik düzeylerini de model içine alan HPM-SAK Model kullanılarak 3 farklı puanlama modeli test edilmiştir.

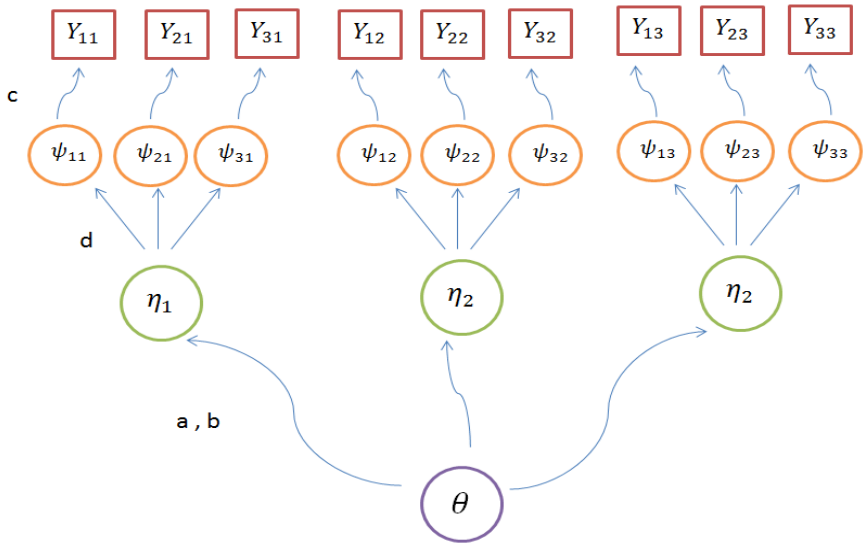
Y_{ji}, j hakeminin i maddesine verdiği puan, η_i, i maddesi, θ gerçek örtük özellik düzeyi, a ve b IRT parametreleri, τ ve φ SDT parametreleri olmak üzere modele ait ilk aşama Şekil 6. ile gösterilmiştir.



Şekil 6. HPM-SAK 1.Model

Bu yaklaşıma göre hakemin doğru kategorileri belirleyebilme becerisi, hakemin puanlama eğilimlerinin bir göstergesidir. Şekil 6. 'da görüldüğü gibi ilk aşamada SAK Model kullanılarak hakem algısına ait parametrelerin (τ, φ) örtük özellik üzerindeki etkisi incelenmiştir.

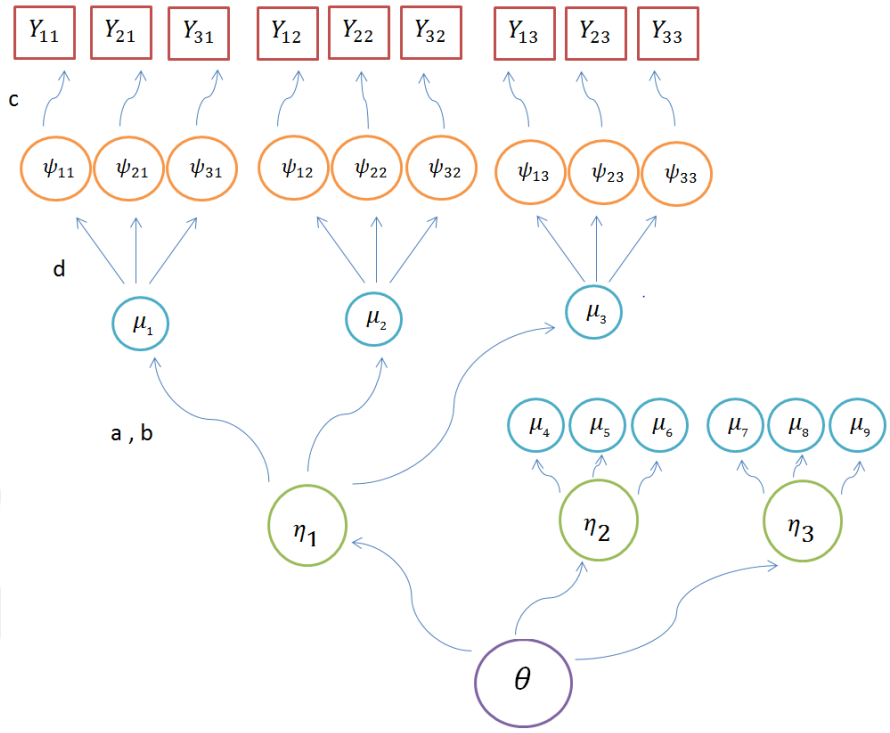
Hiyerarşik puanlayıcı modelinin ikinci aşamasında ise örtük kategoriler, aynı IRT modellerindeki gibi performansların özelliğe sahip olma düzeylerine göre sıralanmasında kullanılır (Richard J Patz et al., 2002). ψ_{ji}, j hakeminin, i maddesi için algısını gösteren örtük özellik olmak üzere, Şekil 7.'de ikinci aşama görselleştirilmiştir.



Şekil 7. HPM-SAK 2.Model

İkinci aşamada HPM-SAT Model kullanılarak madde özelliklerini tanımlayan kategoriler, maddenin örtük sınıfının bir göstergesi olarak kabul edilip, hakem algısı (ψ) da bir örtük özellik olarak ele alınarak incelenmiştir. Şekil 7'de görüldüğü gibi HPM-SAT Modelde örtük özellikle madde arasında doğrusal olmayan, maddeyle puanlayıcı algısı arasında doğrusal, puanlayıcı algısıyla puanlayıcının özlenen puanı arasında doğrusal olmayan ilişkiler bulunmaktadır (DeCarlo, 2011).

Maddeler μ_i, madde niteliğini belirleyen örtük özellik η_i olmak üzere Şekil 8.'de üçüncü aşama görselleştirilmiştir.



Şekil 8. HPM-SAK 3. Model

Üçüncü aşamada ise maddeleri (μ_i), madde niteliğini belirleyen örtük özelliğin (η_i) bir göstergesi olarak kabul edildiği ve maddeye ilişkin kategorik örtük düzeyin belirlendiği hiyerarşik model kurulmuştur.

Bu üç model yukarıda verilen R kodları ile sırasıyla incelenmiş ve modellerin hangisinin daha iyi olduğunu belirlemek için aşağıdaki bilgi kriterleri kullanılmıştır.

AIC (Akaike Information Criterion): Farklı modellerin kıyaslanmasında ve modeller arasından en uygununu seçmek için kullanılmaktadır. $AIC = -2log(L) + 2k$ şeklinde hesaplanan AIC değerinin en küçük olduğu model uygun model olarak yorumlanır.

AICc (Corrected AIC): Küçük örneklemlerde daha iyi sonuçlar elde etmek için AIC kriterinin düzeltilmiş halidir. Bu değer ise $AICc = AIC + (\frac{2k(k+1)}{n-k-1})$ şeklinde hesaplanır.

BIC (Bayesian Information Criterion): Modeller arasında seçim yapmak için kullanılan bir diğer kritertir. AIC'te olduğu gibi en düşük BIC değerine sahip modeller tercih edilir. $BIC = ln(n)k - 2\ln(L)$ şeklinde hesaplanır. Kısmen, olasılık fonksiyonlarına dayanır ve

Akaike bilgi ölçütü (AIC) ile yakından ilişkilidir ancak BIC'de oluşan hata değeri genelde AIC'den daha büyüktür.

CAIC (Consistent Akaike Information Criteria) AIC'in geliştirilmiş ve genişletilmiş daha tutarlı şeklidir. $CAIC = -2logL(\theta_k) + m(k)[\log(n) + 1]$ şeklinde hesaplanır. AIC ve BIC'e benzer şekilde en düşük değere sahip modelin daha iyi olduğu şeklinde yorumlanır.

R kodları ile yapılan analiz sonuçları ve yukarıda verilen bilgi kriterlerine göre elde edilen bulgular bir sonraki bölümde verilmiştir.

3. BULGULAR VE YORUM

Araştırmanın bu bölümde problem ve alt problemlere ilişkin analiz sonuçlarından edilen bulgular, bu bulgulara ilişkin tablo ve yorumlar yer almaktadır.

3.1. Birinci Alt Probleme Ait Bulgular

Test geliştirme sürecinde hakem yargılarına göre madde seçimi aşamasında kurulan HPM modelin uyum düzeyi nedir?

Bu alt probleme yönelik kurulan Model 1'e ait AIC (Akaike information criterion), AICc (AIC corrected), BIC (Bayesian information criterion) ve CAIC (Consistent Akaike Information Criteria) uyum indeksleri hesaplanmış ve aşağıdaki Tablo 10. 'da listelenmiştir

Tablo 10. Model 1'e Ait Uyum İndeksleri

Uyum İndeksleri	Değer	Standart Hata
AIC	26156,66	122
AICc	26191,36	156,7
BIC	26378,38	343,72
CAIC	26439,38	404,72

Tablo 10.'da görüldüğü gibi Model 1'e ait AIC değeri 26156,66 ve bu değere ait standart hata 122, AICc değeri 26191,36 ve bu değere ait standart hata 156,7, BIC değeri 26378,38 ve bu değere ait standart hata 343,72 ve CAIC değeri 26439,38 ve bu değere ait standart hata 404,72'dir. Ayrıca bu modelin hesapladığı EAP (Expected A Posteriori) puanlarının güvenirlik katsayısı Tablo 11.'de verilmiştir.

Tablo 11. Model 1'in Güvenirlik Katsayısı

	Değer
EAP Güvenirlik Katsayısı	0,842

Tablo 11.'de görüldüğü gibi Model 1 için güvenirlik katsayısı 0,842 bulunmuştur. Bu modelin oldukça güvenilir olduğunu göstermektedir.

Model 1 için toplam değerlendirme sayısı (N), maddelerin puan ortalaması (M), 4 kriterli yanıtların kesme noktaları olan c değerleri (Tau.Cat1, Tau.Cat2, Tau.Cat3, Tau.Cat4), maddeleri doğru yanıtlamak için kestirilen yetenek ortalaması (latM) ve yetenek ortalamasının standart sapması (latSS) Tablo 12.'de verilmiştir.

Tablo 12. Model 1 İçin Madde Analizleri

	Model 1								
Madde	N	M	Tau.Cat1	Tau.Cat2	Tau.Cat3	Tau.Cat4	a	latM	latSS
k1	688	3,190	0,474	0,638	0,660	1,753	1	1,683	1,651
k2	688	2,968	-0,663	-0,530	0,243	1,588	1	1,717	1,488
k3	688	3,105	-0,496	-0,165	0,396	1,341	1	1,744	1,565
k4	688	2,824	-0,068	0,112	0,484	2,721	1	1,574	1,470
k5	688	2,413	1,994	10,692	3,581	5,012	1	1,037	1,593
k6	688	3,128	0,534	0,856	0,823	2,034	1	1,632	1,645
k7	688	3,012	-1,258	0,277	0,059	1,303	1	1,730	1,503
k8	688	2,762	-0,952	-0,352	0,394	2,206	1	1,618	1,413
k9	688	2,903	-0,126	0,409	0,823	2,302	1	1,590	1,543
k10	688	3,015	0,698	0,526	1,069	2,305	1	1,586	1,614
k11	688	3,029	1,530	1,316	2,185	2,775	1	1,457	1,694
k12	688	3,010	0,038	0,625	1,133	2,149	1	1,594	1,599
k13	688	2,985	0,335	0,584	0,651	2,482	1	1,585	1,560
k14	688	3,045	-0,196	0,313	0,832	1,830	1	1,654	1,585
k15	688	2,975	0,763	0,697	1,340	2,580	1	1,530	1,609

Tablo 12.'de görüldüğü gibi Model 1'de analize edilen 688 değerlendirme bulunmaktadır. Madde ortalamaları 2,413 ile 3,19 arasında değişmektedir. Kesme kriterleri Tau.Cat1 -1,258 ile 1,994 arasında, Tau.Cat2 -0,53 ile 10,692 arasında, TauCat3 0,059 ile 3,581 ve TauCat4 1,303 ile 5,012 arasında değişmektedir. Ayırt edicilik parametresi olan a değeri sabit tutulduğu için her maddede a=1 verilmiştir. Madde puan dağılımından elde edilen yetenek ortalaması 1,037 ile 1,744 arasında ve yetenek ortalamasının standart sapması 1,413 ile 1,694 arasında değişmektedir.

Tablo 12.'de verilen k5 maddesinin (Soru kültürel, coğrafi ve etnik yanlılık içermiyor.) hakemler tarafından yüksek ortalama ile puanlandığından ve varyansın oldukça düşük olmasından dolayı ilk modelde uyumlu çalışmadığı gözlenmektedir.

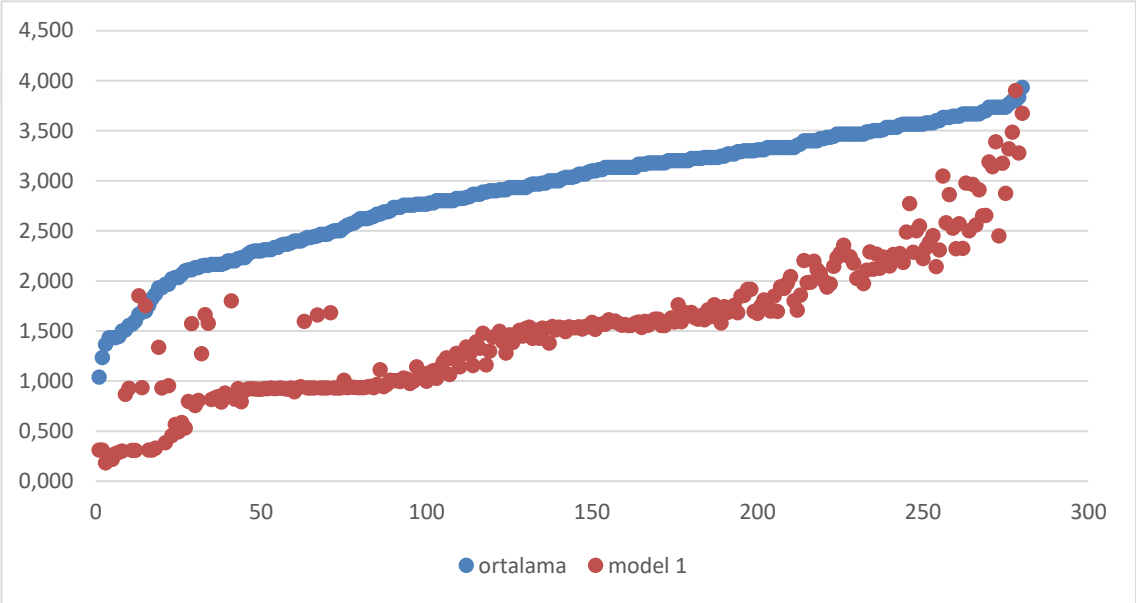
Model 1 için beklenen sonsal yetenek kestirimi (EAP) ve bu kestirimin standart hatası hesaplanmış ve bu değerler Tablo 13.'te verilmiştir.

Tablo 13. Model 1 İçin Beklenen Sonsal Yetenek Kestirimi ve Standart Hata

EAP	SE.EAP
1,541	0,251

Tablo 13.'de görüldüğü gibi Model 1 için yapılan beklenen sonsal yetenek kestirimi (EAP) değeri 1,541 ve bu kestirimin standart hatası 0,251 olarak hesaplanmıştır.

Model 1 ile kestirilen puan dağılımı ve gözlenen puan dağılımı arasındaki ilişki Şekil 9.'da verilmiştir.



Şekil 9. Model 1 İle Kestirilen Puanlar ve Gözlenen Puanlar

Şekil 9.'da verilen mavi noktalar maddelere ait hakem değerlendirmelerinin gözlenen puan ortalamaları kırmızı noktalar ise Model 1 çalıştıktan sonra kestirilen madde puanlarını göstermektedir. Genelde puanlar arasında yüksek bir uyum olduğu görülmektedir ancak model ile kestirilen madde puanları, gözlenen puanlara göre daha düşük ortalamalarda yer almaktadır. Bu durum hakemlerin gözlenen puanlarının Model 1 tarafından daha düşük puanlara doğru düzeltildiğini göstermektedir.

Model 1 ile kestirilen puanlar ile hakem değerlendirmelerinden elde edilen gözlenen puanlar arasındaki Pearson korelasyon katsayısı ve Spearman korelasyon katsayısı hesaplanmış Tablo1 4.'de verilmiştir.

Tablo 14. Model 1 İçin Korelasyon Katsayıları

Pearson Korelasyon katsayısı	0,867**
Spearman Korelasyon Katsayısı	0,935**

**. korelasyon 0,01 düzeyinde anlamlı.

Tablo 14.'de görüldüğü gibi Pearson Korelasyon katsayısı 0,867 olarak hesaplanmıştır. Bu değer gözlenen ve kestirilen puanlar arasında yüksek korelasyon olduğunu gösterir. Pearson korelasyonu puanların farklarının kareleri üzerinden yapılan hesaplamalara dayandığı için gözlenen ve kestirilen puanların birbirine yakın olduğunu göstermektedir. Spearman korelasyon katsayısı 0,935 olarak hesaplanmıştır. Spearman korelasyonu sıralamalar arasındaki farklar üzerinden yapılan hesaplamalara dayandığı için bu değer gözlenen ve kestirilen puanların sıralamalarının çok büyük ölçüde aynı kaldığını göstermektedir.

3.2. İkinci Alt Probleme Ait Bulgular

Test geliştirme sürecinde hakem yargılarına göre madde seçimi aşamasında kurulan 2 düzeyli HPM-SAK uyum düzeyi nedir?

Bu alt probleme yönelik kurulan Model 2 ait. AIC, AICc, BIC ve CAIC uyum indeksleri hesaplanmış ve aşağıdaki Tablo 15. 'da listelenmiştir

Tablo 15. Model 2'ye Ait Uyum İndeksleri

Uyum İndeksi	Değer	Standart Hata
AIC	25816,03	150
AICc	25871,91	205,88
BIC	26088,64	422,61
CAIC	26163,64	497,61

Tablo 15.'da görüldüğü gibi Model 2'ye ait AIC değeri 25816,03 ve bu değere ait standart hata 150, AICc değeri 25871,91ve bu değere ait standart hata 205,88, BIC değeri 26088,64 ve bu değere ait standart hata 422,61 ve CAIC değeri 26163,64 ve bu değere ait standart hata 497,61'dir. Ayrıca bu modelin hesapladığı EAP puanlarının güvenirlik katsayısı Tablo 16.'de verilmiştir.

Tablo 16. Model 2'nin Güvenirlik Katsayısı

	Değer
EAP Güvenirlik Katsayısı	0,900

Tablo 16.'de görüldüğü gibi Model 2 için güvenirlik katsayısı 0,9 bulunmuştur. Bu modelin oldukça güvenilir olduğunu göstermektedir.

Model 2 için toplam değerlendirme sayısı (N), maddelerin puan ortalaması (M), 4 kriterli yanıtların kesme noktaları olan c değerleri (Tau.Cat1, Tau.Cat2, Tau.Cat3, Tau.Cat4), maddeleri doğru yanıtlamak için kestirilen yetenek ortalaması (latM) ve yetenek ortalamasının standart sapması (latSS) Tablo 17.'de verilmiştir.

Tablo 17. Model 2 İçin Madde Analizleri

	Model 2								
item	N	M	Tau.Cat1	Tau.Cat2	Tau.Cat3	Tau.Cat4	a	latM	latSS
k1	688	3,190	0,349	0,596	1,103	2,249	1,092	1,593	1,614
k2	688	2,968	-0,352	-0,056	0,864	2,423	1,104	1,585	1,491
k3	688	3,105	-0,526	0,021	0,745	2,138	1,166	1,634	1,530
k4	688	2,824	0,000	0,255	1,126	3,199	1,108	1,483	1,465
k5	688	2,413	2,074	3,059	2,955	3,538	0,714	1,055	1,583
k6	688	3,128	0,762	0,836	1,022	1,961	0,858	1,580	1,621
k7	688	3,012	0,876	1,401	1,508	1,641	0,566	1,481	1,631
k8	688	2,762	0,442	1,022	2,020	3,928	1,133	1,342	1,498
k9	688	2,903	0,106	0,552	1,257	2,748	1,044	1,502	1,528
k10	688	3,015	0,555	0,764	1,637	2,945	1,100	1,474	1,580
k11	688	3,029	1,329	1,590	2,219	3,080	0,968	1,362	1,639
k12	688	3,010	0,727	1,179	1,861	3,082	1,059	1,423	1,597
k13	688	2,985	0,286	0,683	1,451	3,488	1,262	1,468	1,540
k14	688	3,045	0,335	0,806	1,524	2,615	1,021	1,495	1,579
k15	688	2,975	0,588	0,789	1,671	3,057	1,079	1,450	1,566

Tablo 17.'de görüldüğü gibi Model 2'de analize edilen 688 değerlendirme bulunmaktadır. Madde ortalamaları 2,413 ile 3,19 arasında değişmektedir. Kesme kriterleri Tau.Cat1 -1,258 ile 1,994 arasında, Tau.Cat2 -0,53 ile 10,692 arasında, TauCat3 0,059 ile 3,581 ve TauCat4 1,303 ile 5,012 arasında değişmektedir. Ayırt edicilik parametresi olan a değeri bu modelde sabit değildir ve 0,566 ile 1,262 arasında değişmektedir. Madde puan

dağılımından elde edilen yetenek ortalaması 1,037 ile 1,744 arasında ve yetenek ortalamasının standart sapması 1,413 ile 1,694 arasında değişmektedir.

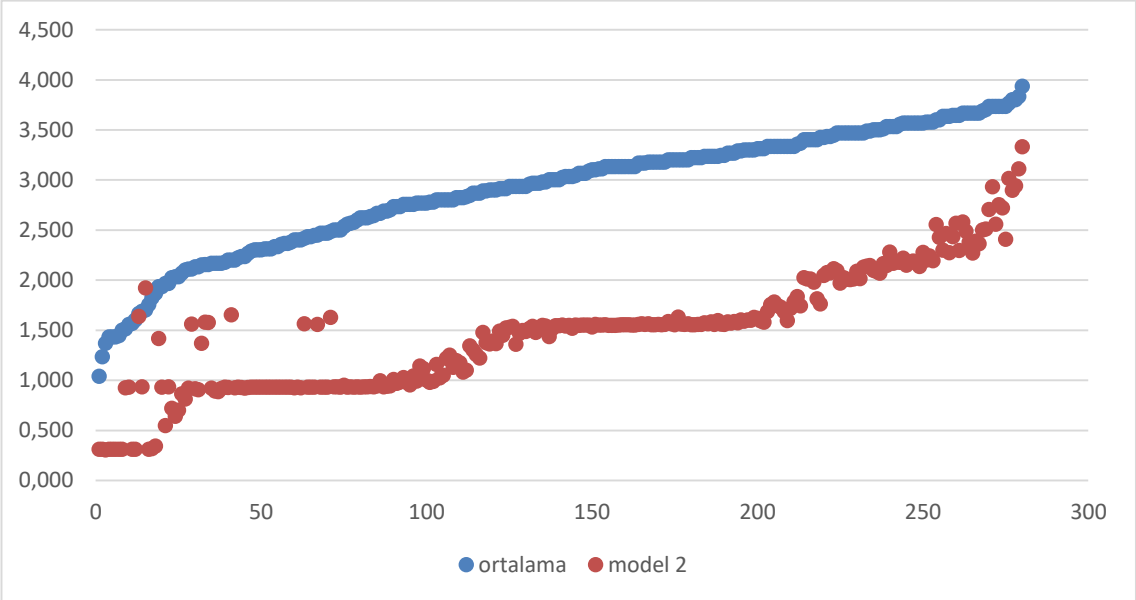
Model 2 için beklenen sonsal yetenek kestirimi (EAP) ve bu kestirimin standart hatası hesaplanmış ve bu değerler Tablo 18.'te verilmiştir.

Tablo 18. Model 2 İçin Beklenen Sonsal Yetenek Kestirimi ve Standart Hata

EAP	SE.EAP
1,485	0,157

Tablo 18.'de görüldüğü gibi Model 2 için yapılan beklenen sonsal yetenek kestirimi (EAP) değeri 1,485 ve bu kestirimin standart hatası 0,157 olarak hesaplanmıştır.

Model 2 ile kestirilen puan dağılımı ve gözlenen puan dağılımı arasındaki ilişki Şekil 10.'da verilmiştir.



Şekil 10. Model 2 İle Kestirilen Puanlar ve Gözlenen Puanlar

Şekil 10.'da verilen mavi noktalar maddelere ait hakem değerlendirmelerinin gözlenen puan ortalamaları kırmızı noktalar ise Model 2 çalıştıktan sonra kestirilen madde puanlarını göstermektedir. Genelde puanlar arasında yüksek bir uyum olduğu görülmektedir ancak model ile kestirilen madde puanları, gözlenen puanlara göre daha düşük ortalamalarda yer almaktadır. Bu durum hakemlerin gözlenen puanlarının Model 2 tarafından daha düşük puanlara doğru düzeltildiğini göstermektedir.

Model 2 ile kestirilen puanlar ile hakem değerlendirmelerinden elde edilen gözlenen puanlar arasındaki Pearson korelasyon katsayısı ve Spearman korelasyon katsayısı hesaplanmış Tablo 19.'de verilmiştir.

Tablo 19. Model 2 İçin Korelasyon Katsayıları

Pearson Korelas yon katsayısı	0,876**
Spearman Korelasyon Katsayısı	0,929**

**. korelasyon 0,01 düzeyinde anlamlı.

Tablo 19.'de görüldüğü gibi Pearson Korelasyon katsayısı 0,876 olarak hesaplanmıştır. Bu değer gözlenen ve kestirilen puanlar arasında yüksek korelasyon olduğunu gösterir. Pearson korelasyonu puanların farklarının kareleri üzerinden yapılan hesaplamalara dayandığı için gözlenen ve kestirilen puanların birbirine yakın olduğunu göstermektedir. Spearman korelasyon katsayısı 0,929 olarak hesaplanmıştır. Spearman korelasyonu sıralamalar arasındaki farklar üzerinden yapılan hesaplamalara dayandığı için bu değer gözlenen ve kestirilen puanların sıralamalarının çok büyük ölçüde aynı kaldığını göstermektedir.

3.3. Üçünü Alt Probleme Ait Bulgular

Test geliştirme sürecinde hakem yargılarına göre madde seçimi aşamasında kurulan 3 düzeyli HPM-SAK uyum düzeyi nedir?

Bu alt probleme yönelik kurulan Model 3'e ait. AIC, AICc, BIC ve CAIC uyum indeksleri hesaplanmış ve aşağıdaki Tablo 20.'da listelenmiştir

Tablo 20. Model 3'e Ait Uyum İndeksleri

Uyum İndeksi	Değer	Standart Hata
AIC	25983,79	132
AICc	26025,31	173,52
BIC	26223,68	371,9
CAIC	26289,68	437,9

Tablo 20.'da görüldüğü gibi Model 3'e ait AIC değeri 25983,79 ve bu değere ait standart hata 132, AICc değeri 26025,31 ve bu değere ait standart hata 173,52, BIC değeri 26223,68 ve bu değere ait standart hata 371,9 ve CAIC değeri 26289,68 ve bu değere ait standart hata 437,9'dir. Ayrıca bu modelin hesapladığı EAP puanlarının güvenirlik katsayısı Tablo 21.'de verilmiştir.

Tablo 21. Model 3'ün Güvenirlik Katsayısı

	Değer
EAP Güvenirlik Katsayısı	0,873

Tablo 21.'de görüldüğü gibi Model 3 için güvenirlik katsayısı 0,873 bulunmuştur. Bu modelin oldukça güvenilir olduğunu göstermektedir.

Model 3 için toplam değerlendirme sayısı(N), maddelerin puan ortalaması(M), 4 kriterli yanıtların kesme noktaları olan c değerleri (Tau.Cat1, Tau.Cat2, Tau.Cat3, Tau.Cat4), maddeleri doğru yanıtlamak için kestirilen yetenek ortalaması (latM) ve yetenek ortalamasının standart sapması (latSS) Tablo 22.'de verilmiştir.

Tablo 22. Model 3 İçin Madde Analizleri

	Model 3								
item	N	M	Tau.Cat1	Tau.Cat2	Tau.Cat3	Tau.Cat4	a	latM	latSS
k1	688	3,190	1,089	-1,510	-0,643	0,726	1	1,924	1,530
k2	688	2,968	2,095	-0,812	0,550	2,416	1	1,660	1,503
k3	688	3,105	-0,497	-2,446	-1,222	0,174	1	1,998	1,403
k4	688	2,824	1,108	-1,281	-0,068	2,619	1	1,690	1,399
k5	688	2,413	-1,792	8,913	0,965	3,207	1	1,447	1,338
k6	688	3,128	-0,424	-1,659	-1,003	0,723	1	1,923	1,464
k7	688	3,012	1,635	-1,540	-0,318	1,341	1	1,836	1,480
k8	688	2,762	0,667	-0,892	0,693	2,874	1	1,602	1,428
k9	688	2,903	0,516	-1,015	0,275	2,147	1	1,700	1,473
k10	688	3,015	2,425	-0,729	0,676	2,140	1	1,684	1,549
k11	688	3,029	0,328	-0,437	0,958	2,032	1	1,668	1,567
k12	688	3,010	-0,237	-1,271	0,045	1,448	1	1,784	1,481
k13	688	2,985	0,092	-1,593	-0,666	1,550	1	1,823	1,421
k14	688	3,045	1,701	-0,872	0,652	1,903	1	1,715	1,548
k15	688	2,975	0,325	-1,046	0,160	1,912	1	1,730	1,485

Tablo 22.'de görüldüğü gibi Model 3'de analize edilen 688 değerlendirme bulunmaktadır. Madde ortalamaları 2,413 ile 3,19 arasında değişmektedir. Kesme kriterleri Tau.Cat1 -1,258 ile 1,994 arasında, Tau.Cat2 -0,53 ile 10,692 arasında, TauCat3 0,059 ile 3,581 ve TauCat4 1,303 ile 5,012 arasında değişmektedir. Ayırt edicilik parametresi olan a değeri sabit tutulduğu için her maddede a=1 verilmiştir. Madde puan dağılımından elde edilen yetenek ortalaması 1,037 ile 1,744 arasında ve yetenek ortalamasının standart sapması 1,413 ile 1,694 arasında değişmektedir.

Tablo 22.'de verilen k5 maddesinin (Soru kültürel, coğrafi ve etnik yanlılık içermiyor.) hakemler tarafından yüksek ortalama ile puanlandığından ve varyansın oldukça düşük olmasından dolayı ilk modelde uyumlu çalışmadığı gözlenmektedir.

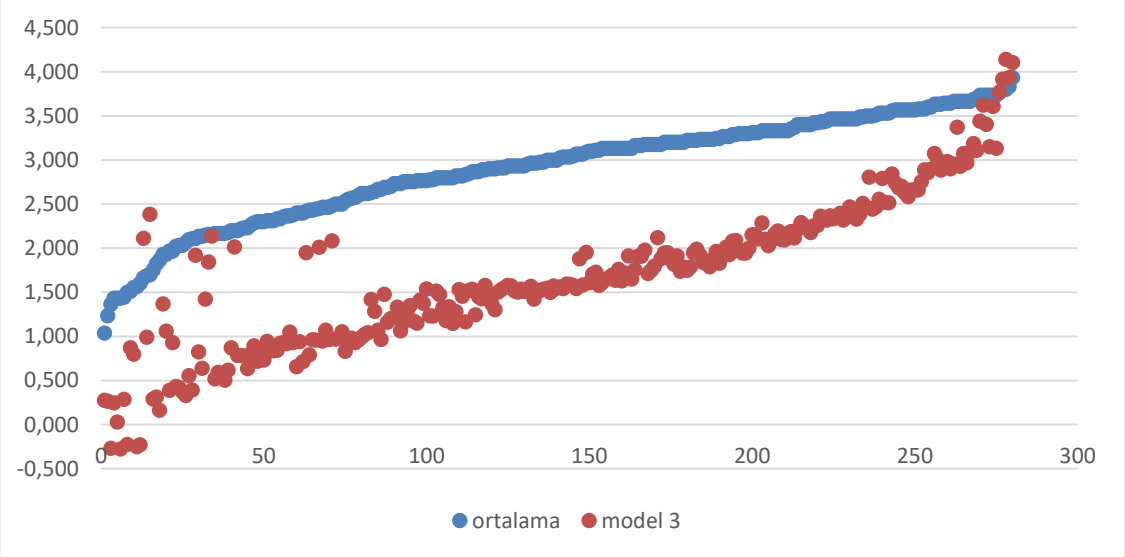
Model 3 için beklenen sonsal yetenek kestirimi (EAP) ve bu kestirimin standart hatası hesaplanmış ve bu değerler Tablo 23.'te verilmiştir.

Tablo 23. Model 3 İçin Beklenen Sonsal Yetenek Kestirimi ve Standart Hata

EAP	SE.EAP
1,691	0,292

Tablo 23.'te görüldüğü gibi Model 3 için yapılan beklenen sonsal yetenek kestirimi (EAP) değeri 1,691 ve bu kestirimin standart hatası 0,292 olarak hesaplanmıştır.

Model 3 ile kestirilen puan dağılımı ve gözlenen puan dağılımı arasındaki ilişki Şekil 11.'de verilmiştir.



Şekil 11. Model 3 İle Kestirilen Puanlar ve Gözlenen Puanlar

Şekil 11.'da verilen mavi noktalar maddelere ait hakem değerlendirmelerinin gözlenen puan ortalamaları kırmızı noktalar ise Model 3 çalıştıktan sonra kestirilen madde puanlarını göstermektedir. Genelde puanlar arasında yüksek bir uyum olduğu görülmektedir ancak model ile kestirilen madde puanları, gözlenen puanlara göre daha düşük ortalamalarda yer almaktadır. Bu durum hakemlerin gözlenen puanlarının Model 3 tarafından daha düşük puanlara doğru düzeltildiğini göstermektedir.

Model 3 ile kestirilen puanlar ile hakem değerlendirmelerinden elde edilen gözlenen puanlar arasındaki Pearson korelasyon katsayısı ve Spearman korelasyon katsayısı hesaplanmış Tablo 24.'de verilmiştir.

Tablo 24. Model 3 İçin Korelasyon Katsayıları

Pearson Korelas yon katsayısı	0,877**
Spearman Korelasyon Katsayısı	0,924**

**. korelasyon 0,01 düzeyinde anlamlı.

Tablo 24.'de görüldüğü gibi Pearson Korelasyon katsayısı 0,877 olarak hesaplanmıştır. Bu değer gözlenen ve kestirilen puanlar arasında yüksek korelasyon olduğunu gösterir. Pearson korelasyonu puanların farklarının kareleri üzerinden yapılan hesaplamalara dayandığı için gözlenen ve kestirilen puanların birbirine yakın olduğunu göstermektedir. Spearman korelasyon katsayısı 0,924 olarak hesaplanmıştır. Spearman korelasyonu sıralamalar arasındaki farklar üzerinden yapılan hesaplamalara dayandığı için bu değer gözlenen ve kestirilen puanların sıralamalarının çok büyük ölçüde aynı kaldığını göstermektedir.

4. SONUÇ VE ÖNERİLER

Bu bölümde araştırma verileri ile elde edilen analiz sonuçları ve bir sonraki araştırmalar için çeşitli öneriler verilmiştir.

4.1. Sonuçlar

Araştırmada 115K531 No'lu TÜBİTAK projesi kapsamında geliştirilen ve yazarların oluşturduğu maddeleri değerlendirmek için kullanılan puanlama anahtarı ve bu anahtar ile hakemlerin madde değerlendirmeleri kullanılmıştır. Hakem yanlılığının yanı sıra hakem cömertlik ve sertlik parametrelerine de yer veren üç ayrı model denenmiştir.

Model 1'de Y_{ji}, j hakeminin i maddesine verdiği puan, η_i, i maddesi, θ gerçek örtük özellik düzeyi, a ve b IRT parametreleri, τ ve φ SDT parametreleri olmak üzere SAK Model kullanılarak hakem algısına ait parametrelerin (τ, φ) örtük özellik üzerindeki etkisi incelenmiştir.

Model 2'de, HPM-SAT Model kullanılarak madde özelliklerini tanımlayan kategoriler, maddenin örtük sınıfının bir göstergesi olarak kabul edilip, hakem algısı (ψ) da bir örtük özellik olarak ele alınarak incelenmiştir.

Model 3'de maddeler μ_i, madde niteliğini belirleyen örtük özellik η_i olmak üzere, maddeleri (μ_i), madde niteliğini belirleyen örtük özelliğin (η_i) bir göstergesi olarak kabul edildiği ve maddeye ilişkin kategorik örtük düzeyin belirlendiği hiyerarşik model kurulmuştur. Alt problemlere yönelik kurulan üç modele ait hesaplanan AIC, AICc, BIC ve CAIC gibi uyum indeksleri aşağıdaki Tablo 25.'de listelenmiştir

Tablo 25. Üç Modele Ait Uyum İndeksleri

	Model 1		Model 2		Model 3	
	Değer	St.Hata	Değer	St.Hata	Değer	St.Hata
AIC	26156,66	122	25816,03	150	25983,79	132
AICc	26191,36	156,7	25871,91	205,88	26025,31	173,52
BIC	26378,38	343,72	26088,64	422,61	26223,68	371,9
CAIC	26439,38	404,72	26163,64	497,61	26289,68	437,9

Tablo 25.'de verilen üç modelin uyumları incelendiğinde, kurulan 3 ayrı model için verilen tüm bilgi kriterlerine bakıldığında Ayrıca en düşük AIC = 26156,66, AICc = 26191,36, BIC = 26378,38 ve CAIC = 26439,38 değerleri ile 2. Modelin data ile diğer modellere göre daha uyumlu olduğu söylenebilir. Model 3'ün de Model 1'e nazaran daha tüm kriterler için daha iyi uyum verdiği görülmektedir.

Üç modelin güvenirlik katsayıları Tablo 26. 'da verilmiştir.

Tablo 26. Üç Ayrı Modelin Güvenirlik Katsayısı

	Model 1	Model 2	Model 3
EAP Güvenirlik katsayısı	0,842	0,9	0,873

Tablo 26.'da görüldüğü gibi 0,9 EAP Güvenirlik katsayısı ile 2.Model'in diğerlerinden daha güvenilir sonuçlar verdiği elde edilmiştir. Model 3'ün de Model 1'e göre daha güvenilir sonuçlar verdiği görülmektedir.

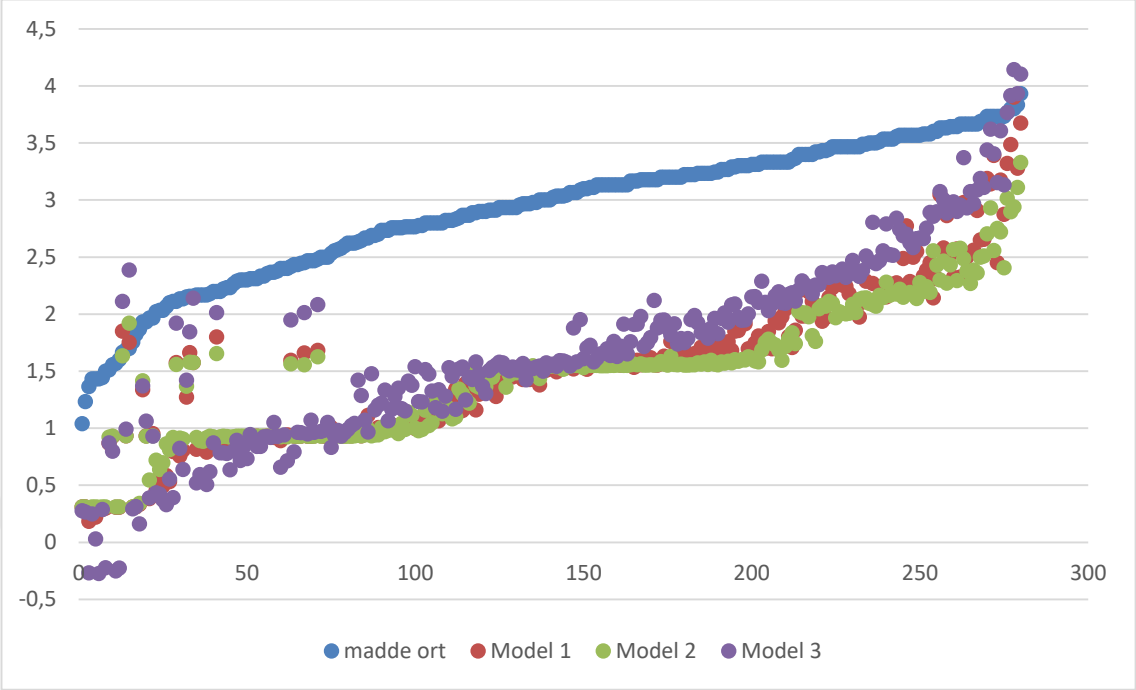
Üç model için beklenen sonsal yetenek kestirimleri (EAP) ve bu kestirimin standart hataları Tablo 27.'de verilmiştir.

Tablo 27. Üç Model İçin Beklenen Sonsal Yetenek Kestirimi ve Standart Hata

	Model 1	Model 2	Model 3
EAP	1,541	1,485	1,691
EAP SH	0,251	0,157	0,292

Tablo 27.'de görüldüğü gibi 3 model için yapılan beklenen sonsal yetenek kestirim (EAP) değeri içinde en yüksek değer Model 3'e aitken bu model için hesaplanan kestirim hatası (EAP SH) da en büyük değere sahiptir. 1,485 beklenen sonsal yetenek kestirime sahip olan 2. Modelin bu kestirime ait hatası 0,157'dir ve diğer modellere göre daha iyi bir hata değerine sahip olduğu görülmüştür.

Üç model ile kestirilen puan dağılımı ve gözlenen puan dağılımı arasındaki ilişki Şekil 1 2.'de verilmiştir.



Şekil 12. Üç Model ile Kestirilen Puanlar ve Gözlenen Puanlar

Şekil 12.'da verilen mavi noktalar maddelere ait hakem değerlendirmelerinin gözlenen puan ortalamaları, kırmızı, yeşil ve mor noktalar ise sırasıyla Model 1, Model 2 ve Model 3 çalıştıktan sonra kestirilen madde puanlarını göstermektedir.

Model 3'ün daha eğrisel bir yetenek dağılım yaptığı, daha düşük ve daha yüksek puanlara çıkarak daha değişken olduğu ve gözlenen puanlara daha çok yaklaştığı görülmektedir. Model 1 ve Model 2 daha sabit dağılımlara sabit olduğu görünmektedir.

Üç model ile kestirilen puanlar ve hakem değerlendirmelerinden elde edilen gözlenen puanlar arasındaki Pearson ve Spearman korelasyon katsayısı hesaplanmış Tablo 28.'de verilmiştir.

Tablo 28. Üç Model İçin Korelasyon Katsayıları

	Model 1	Model 2	Model 3
Pearson Korelas yon katsayısı	0,867**	0,876**	0,877**
Spearman Korelasyon Katsayısı	0,935**	0,929**	0,924**

**. korelasyon 0,01 düzeyinde anlamlı.

Tablo 28.'de görüldüğü gibi Model 3'e ait Pearson Korelasyon katsayısı en yüksek iken, Model 1'e ait Spearman Korelasyon Katsayısı en yüksektir. 0,876 Pearson Korelasyon Katsayısı ve 0,929 Spearman Korelasyon katsayısı ile 2.Model'in gözlenen puanlarla korelasyonlarının diğer modellere oldukça yakın değerler verdiği elde edilmiştir

Tüm bu bilgi kriterleri incelenerek çalışmanın örneklemi için 2. Modelin diğer modellerden daha iyi çalıştığı söylenebilir. Kurulan Model 2'de hakemlerin maddeler için verdiği puanlar soru niteliğinin göstergeleri olarak kabul edilmiştir. Aynı zamanda hakem algısı da modele bir diğer faktör olarak dâhil edilmiştir. Soru sayısının (örneklem büyüklüğü) MTK modelleri için kısmen az olduğu göz önüne alınırsa Model 2'nin Model 3'e göre daha uyumlu çıkmasının istatistiksel olarak normal olduğu söylenebilir.

Hakem katılığının modele dâhil olduğunda parametre sayısının artması kestirimlere ilişkinin hatanın da arttığının bir göstergesi olabilir. Bu durumda daha büyük örneklemlerde Model 3'ün diğer modellere göre data ile daha yüksek uyum sağlayabileceği düşünülmektedir.

Bununla birlikte test geliştirme süreci için en kritik basamaklardan biri olan soru yazma ve değerlendirme sürecini daha geniş bir perspektifle ve bu sürece etki eden daha fazla faktörü hesaplamalara dâhil ederek istatistiksel olarak sınanabilir bir yöntem olarak HPM-SAK Modelin kullanılabileceği görülmektedir.

Ayrıca kullanılan bu model uygulayıcılara yapısal bir karşılaştırma olanağı sunmakla kalmayıp ölçülen özellik, ölçme aracı ve hakemlere ait hem SAK modeline göre hem de MTK modeline göre parametreler vermektedir.

Sadece hakem algılarına dayalı değerlendirmelerde hakem yanlılığı, hakem güvenirliği ve hakem algısı gibi parametreler ile kararlar verilirken HPM-SAK Model aracılığıyla daha açıklayıcı parametrelere ulaşma imkânı da bulunmaktadır.

4.2. Öneriler

- Araştırmada kullanılan modellerin örneklem büyüklüğü, hakem sayısı ve ölçme aracı duyarlılığı konusunda literatürde yeterince çalışma bulunmamaktadır. Araştırmada proje örnekleminde bağlı kaldığından yukarda belirtilen koşullara ilişkin herhangi bir analize yer verilmemiştir. Bu durum göz önüne alındığında bundan sonra yapılacak çalışmalarda örneklem büyüklü, hakem sayısı ve ölçme aracı duyarlılığının etkisi araştırılabilir.
- Bu modellerin etkinliğini belirlemek amacıyla kurgulanacak bir araştırmada soru niteliğine ilişkin pilot uygulama verileri elde edilerek önsel çalışmanın gerçek durumdaki karşılığı belirlenebilir. Bu araştırmada puanlanan soruların tamamının pilot uygulamaya seçilmemesi HPM-SAK Modelle belirlenen sorulara ilişkin önsel bilgilerin pilot uygulama ile karşılaştırılmasına imkan vermemiştir. Sadece bu amaç için özel bir araştırma deseni bundan sonra yapılacak çalışmalar için kurgulanabilir.
- Araştırmada kurulan model hiyerarşik bir yapı taşıdığından örtük özellikler ve göstergelerin tekrar tasarlanmasına olanak sağlamaktadır. Bu anlamda bu modele ilişkin ölçme aracı maddelerinin başka alt boyutların göstergesi olduğu kabul edilerek modele yeni bir düzey daha eklenebilir. Bu çalışmada geliştirilen ölçme aracı tek boyutlu kabul edildiğinden bu konuda herhangi bir analize yer verilmemiştir.

KAYNAKÇA

Agresti, A. (2002). Inference for contingency tables. *Categorical Data Analysis, Second Edition*, 70–114.

Baker, F. B. (2001). *The basics of item response theory*. ERIC.

Banks, W. P. (1970). Signal detection theory and human memory. *Psychological Bulletin*, *74*(2), 81.

Benkhoff, B. (1997). A test of the HRM model: Good for employers and employees. *Human Resource Management Journal*, *7*(4), 44–60.

Bollen, K. (1989). Structural Equations with Latent Variables. New York: John Wiley & Sons, Inc. Wiley, DE (1973) The Identification Problem for Structural Equation Models with Unmeasured Variables in Goldberger AS and Duncan, OD eds. Tructural Equation Models in the Social Sciences-New York: Academic Press.

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, *12*(4), 335–359.

Clogg, C. C., & Manning, W. D. (1996). Assessing Reliability of Categorical Measurements Using Latent Class Models-8.

Coffman, W. E. (1971). Essay examinations. *Educational Measurement*, *2*, 271–302.

Cronbach, L. J. (1989). Construct validation after thirty years. *Intelligence: Measurement, Theory, and Public Policy*, *3*, 147–171.

De Ayala, R. J. (2013). *The theory and practice of item response theory*. Guilford Publications.

Decarlo, L. T. (2002). A Latent Class Extension of Signal Detection Theory , with Applications. *Multivariate Behavioral Research*. https://doi.org/10.1207/S15327906MBR3704

Decarlo, L. T. (2005). A Model of Rater Behavior in Essay Grading Based on Signal Detection Theory. *Journal of Educational Measurement Spring*. https://doi.org/10.1111/j.0022-0655.2005.00004.x

Decarlo, L. T. (2008). Studies of a Latent-Class Signal-Detection Model for Constructed-Response Scoring. *ETS Research Report Series*.

DeCarlo, L. T. (1998). Signal detection theory and generalized linear models. *Psychological Methods*. https://doi.org/10.1037/1082-989X.3.2.186

DeCarlo, L. T. (2003). Source monitoring and multivariate signal detection theory, with a model for selection. *Journal of Mathematical Psychology*.

DeCarlo, L. T. (2008). On a hierarchical rater model for essay grading: Incorporating a latent class signal detection model. In *annual meeting of the National Council on Measurement in Education, New York.*

DeCarlo, L. T. (2010). On the Analysis of Fraction Subtraction Data: The DINA Model, Classification, Latent Class Sizes, and the Q-Matrix. *Applied Psychological Measurement*, *35*(1), 8–26. https://doi.org/10.1177/0146621610377081

DeCarlo, L. T. (2010). Studies of a Latent Class Signal Detection Model for Constructed Response Scoring II: Incomplete and Hierarchical Designs. Research Report. ETS RR-10-08. *Educational Testing Service*.

DeCarlo, L. T. (2011). Signal detection theory with item effects. *Journal of Mathematical Psychology*. https://doi.org/10.1016/j.jmp.2011.01.002

DeCarlo, L. T., Kim, Y., & Johnson, M. S. (2011). A hierarchical rater model for constructed responses, with a signal detection rater model. *Journal of Educational Measurement*, *48*(3), 333–356. https://doi.org/10.1111/j.1745-3984.2011.00143.x

DeCarlo, L. T., & Kim, Y. K. (2009). On scoring constructed response items and multiple choice items: Incorporating signal detection and item response models into a hierarchical rater model. In *annual meeting of the National Council on Measurement in Education, San Diego, CA.*

Donoghue, J. R., & Hombo, C. M. (2000). A comparison of different model assumptions about rater effects. In *Annual Meeting of the National Council on Measurement in Education*.

Downing, S. M., & Haladyna, T. M. (2006). *Handbook of test development.* Lawrence Erlbaum Associates Publishers.

Egan, J. P. (1975). Signal detection theory and {ROC} analysis.

Freeman, E., Heathcote, A., Chalmers, K., & Hockley, W. (2010). Item effects in recognition memory for words. *Journal of Memory and Language*, *62*(1), 1–18.

Gescheider, G. A. (1997). Psychophysics: The Fundamentals. Lawrence Erlbaum Associates. *Mahwah, NJ*.

Green, D. M., & Swets, J. A. (1966). Signal detection theory and psychophysics Wiley.[arLEK, DDD. NAM, MT].

Green, D. M., & Swets, J. A. (1988). Signal Detection Theory and Psychophysics (Peninsula, Los Altos, CA).

Gronlund, N. E. (1982). *Constructing achievement tests*. Prentice-Hall.

Hasselblad, V., & Hedges, L. V. (1995). Meta-analysis of screening and diagnostic tests. *Psychological Bulletin*, *117*(1), 167.

Irvine, S. H., & Kyllonen, P. C. (2013). *Item generation for test development*. Routledge.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement*, *38*(1), 79–93.

Karasar, N. (1998). Bilimsel Araştırma Yöntemi-Kavramlar, İlkeler, Teknikler-, Nobel Yayın Dağıtım, 8. *Basım, Ankara*.

Linacre, J. M. (1989). Many-faceted Rasch measurement. Univ. of Chicago, Dept. of Education.

Macmillan, N. A., & Creelman, C. D. (1991). Detection Theory: A User's Guide

(Cambridge UP, Cambridge).

Macmillan, N. A., & Creelman, C. D. (2005). Detection theory: A user's guide.

Mariano, L. T. (2002). *INFORMATION ACCUMULATION , MODEL SELECTION AND RATER BEHAVIOR IN CONSTRUCTED RESPONSE STUDENT ASSESSMENTS*. *Statistics*.

Mariano, L. T., & Junker, B. W. (2007). Covariates of the Rating Process in Hierarchical Models for Multiple Ratings of Test Items. *Journal of Educational and Behavioral Statistics*. https://doi.org/10.3102/1076998606298033

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174.

McNicol, D. (2005). *A primer of signal detection theory*. Psychology Press.

Morey, R. D., Pratte, M. S., & Rouder, J. N. (2008). Problematic effects of aggregation in z ROC analysis and a hierarchical modeling solution. *Journal of Mathematical Psychology*, *52*(6), 376–388.

Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., & Preuschoff, C. (2009). *TIMSS 2011 Assessment Frameworks.* ERIC.

Muraki, E. (1992). A GENERALIZED PARTIAL CREDIT MODEL: APPLICATION OF AN EM ALGORITHM. *ETS Research Report Series*, *1992*(1), i-30. https://doi.org/10.1002/j.2333-8504.1992.tb01436.x

Murphy, K. R., & Davidshofer, C. O. (1988). Psychological testing. *Principles, and Applications, Englewood Cliffs*.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement*, *5*(2), 189–227.

Patz, R. J. (1996). Markov chain Monte Carlo methods for item response theory models with applications for the National Assessment of Educational Progress. *Unpublished*

Doctoral Dissertation, Carnegie Mellon University, Pittsburgh PA.

Patz, R. J., & Junker, B. W. (1997). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*.

Patz, R. J., Junker, B. W., Johnson, M. S., & Mariano, L. T. (2002). The Hierarchical Rater Model for Rated Test Items and Its Application to Large-Scale Educational Assessment Data. *Journal of Educational and Behavioral Statistics*, *27*(4), 341–384. https://doi.org/10.3102/10769986027004341

Pratte, M. S., Rouder, J. N., & Morey, R. D. (2010). Separating mnemonic process from participant and item effects in the assessment of ROC asymmetries. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(1), 224.

Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review*, *12*(4), 573–604.

Rouder, J. N., Lu, J., Sun, D., Speckman, P., Morey, R., & Naveh-Benjamin, M. (2007). Signal detection models with random participant and item effects. *Psychometrika*, *72*(4), 621–642.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*.

Singh, S., & Churchill, G. A. (1986). Using the theory of signal detection to improve ad recognition testing.

Stein, M. K., & Smith, M. S. (1998). Mathematical tasks as a framework for reflection: From research to practice. *Mathematics Teaching in the Middle School*, *3*(4), 268–275.

Swets, J. A. (1986). Form of empirical ROCs in discrimination and diagnostic tasks: Implications for theory and measurement of performance. *Psychological Bulletin*, *99*(2), 181.

Swets, J. A. (2014). *Signal detection theory and ROC analysis in psychology and diagnostics: Collected papers*. Psychology Press.

Tanner Jr, W. P., & Swets, J. A. (1954). A decision-making theory of visual detection. *Psychological Review*, *61*(6), 401.

Wickens, T. D. (2002). *Elementary signal detection theory*. Oxford University Press, USA.

Wilson, M., & Hoskens, M. (2001). The rater bundle model. *Journal of Educational and Behavioral Statistics*, *26*(3), 283–306.

ÖZGEÇMİŞ

03.01.1990 tarihinde Ankara'da doğdu. İlköğretimini ve orta öğretimini Burdur Gazi İlköğretim Okulu'nda, lise öğretimini Burdur Anadolu Lisesi'nde tamamladı. 2014

yılında Ege Üniversitesi, Fen Fakültesi, Matematik Bölümü, Teorik Ağırlıklı Matematik Lisans öğretim programından mezun oldu. 2014 yılında Ege Üniversitesi, Fen Bilimleri Enstitüsü, Matematik Bölümü, Analiz ve Fonksiyonlar Teorisi Anabilim Dalı'nda tezli yüksek lisans programına başladı ve 2016 yılında bu programı tamamladı. 2015 yılında Ege Üniversitesi, Sosyal Bilimler Enstitüsü, Eğitim Bilimleri Anabilim Dalı, Eğitimde Ölçme ve Değerlendirme Bilim Dalı'nda yüksek lisans programına başladı ve 2017 yılında bu programı tamamladı. Aynı yıl Hacettepe Üniversitesi, Eğitim Bilimleri Enstitüsü, Eğitim Bilimleri Anabilim Dalı, Eğitimde Ölçme ve Değerlendirme Bilim Dalı'nda doktora programına katılmaya hak kazandı.

ÖZET

Test Geliştirme Sürecinde Madde Seçimi İçin Hiyerarşik Puanlayıcı Model: Sinyal Algılama Puanlayıcı Modelinin Uyarlanması

Bu araştırma kısa cevaplı, açık uçlu ve boşluk doldurma gibi hakem puanlamalarına dayanan soruların hakem yanlılığından arındırılarak puanlanmasına yönelik geliştirilmiş olan HPM-SAK modelinin, test geliştirme sürecinin madde seçimin aşamasında soru kalitesinin belirlenmesi amacıyla kullanılmasına dayanmaktadır. Üç düzey olarak tasarlanan bu hiyerarşik modelde hakem niteliği ve hakem katılığı iki ayrı faktör olarak ele alınmaktadır. Bu iki faktör belirlenirken ilk aşamada Madde Tepki Kuramı ve ikinci aşamada Sinyal Algılama Teorisi parametreleri kullanılmaktadır. Model bütün olarak gözlenen hakem puanlarından yola çıkarak bir sorunun öğrenci yeteneğini kestirmedeki düzeyini hakem niteliği ve hakem katılığı etkilerini kontrol ederek tekrar hesaplamaktadır. Araştırmada bu yaklaşım test geliştirme aşamasında hakem kararlarına dayalı olarak soru niteliğini belirleme sürecinde hakem etkisinin kontrol edilebilmesi amacıyla kullanılmıştır.

Araştırma 115K531 No'lu TÜBİTAK projesinin akademik çıktılarının bir boyutunu oluşturmaktadır. Proje kapsamında geliştirilen 280 soru 28 hakem tarafından 15 maddeden oluşan puanlama ölçeği ile değerlendirilmiştir. Bu süreçte geliştirilen her soru üç hakem tarafından incelenmiştir. Araştırmada hakem incelemeleri sonucunda sorular için verilen puanlar HPM-SAK model ile analiz edilmiştir. HPM-SAK model üç düzeyli hiyerarşik bir yapı taşımaktadır. Bu nedenle bu düzeylere göre önerilen üç model ayrı ayrı analiz edilmiş ve hangi modelin data ile daha uyumlu olduğu belirlenmeye çalışılmıştır.

Analiz sonuçlarına göre HPM-SAK için ikinci modelin diğer modellere göre daha iyi uyum gösterdiği görülmüştür. En düşük uyum düzeyi birinci düzey modelde gözlenmiştir. Puanlama güvenirlikleri incelendiğinde ikinci model indeksinin diğer modellerinden daha yüksek olduğu belirlenmiştir. Bu durum HPM-SAK modelin soru niteliğini belirleme sürecinde alternatif bir güvenirlik yaklaşımı olarak kullanılabileceğinin bir göstergesidir.

Anahtar Sözcükler: SAK Model, HPM Model, HPM-SAK Model, MTK

ABSTRACT

The Hierarchical Rater Model in Item Selection of Test Development Process: Adopting of Signal Detection Rater Model

This research is based on the use of the HRM-SDT model developed for the scoring of questions based on referee scoring, such as short answer, open ended and gap filling, in order to determine the quality of the item during the item selection phase. In this hierarchical model, which is designed as three levels, the referee is modeled as two factors of quality and referee participation. When these two factors are determined, the first stage uses the Matter Response Theory and the second uses the Signal Detection Theory parameters. The model recalculates the level of a student's ability to judge the student's ability by controlling the referee's qualifications and referee involvement effects, based on the overall referee points. This approach has been used in the research in order to control the referee effect in the process of determining the substance quality based on the referee's decisions during the test development stage.

This research constitutes a dimension of the academic outputs of 115K531 TUBITAK project. 280 questions developed within the scope of the project were evaluated by 28 referees with a scoring scale consisting of 15 items. Each question developed in this process was examined by three referees. As a result of the referee examinations in the research, the scores given for the questions were analyzed with the HRM-SDT model. The HRM-SDT model has a three-level hierarchical structure. For this reason, the three models proposed according to these levels were analyzed separately and it was tried to determine which model is more compatible with the data.

According to the analysis results, the second model for HRM-SDT is seen to be better than the other models. The lowest level of compliance was observed in the first level model. When the reliability of rating was examined, it was determined that the second model index was higher than the others. This is an indication that the HRM-SDT model can be used as an alternative reliability approach in the process of determining the substance quality.

Key Words: SDT Model, HRM Model, HRM-SDT Model, IRT

EKLER

EK1. Maddelere Atanan Hakem Listesi

1	5199	21	hakem17	hakem33	hakem18	48	9692	21	hakem22	hakem34	hakem26	95	6963	55	hakem12	hakem11	hakem26
2	6936	21	hakem33	hakem13	hakem30	49	5106	22	hakem35	hakem26	hakem24	96	5057	55	hakem29	hakem30	hakem15
3	6381	21	hakem12	hakem28	hakem16	50	1979	30	hakem11	hakem15	hakem29	97	9892	55	hakem12	hakem28	hakem34
4	4420	21	hakem38	hakem23	hakem32	51	8492	31	hakem37	hakem20	hakem18	98	6470	55	hakem29	hakem34	hakem33
5	7104	21	hakem36	hakem21	hakem24	52	5856	47	hakem31	hakem35	hakem21	99	6096	55	hakem37	hakem36	hakem34
6	7943	21	hakem37	hakem25	hakem32	53	1289	48	hakem35	hakem32	hakem14	100	8642	55	hakem28	hakem30	hakem35
7	1626	22	hakem14	hakem32	hakem18	54	7875	54	hakem26	hakem35	hakem31	101	4945	61	hakem13	hakem32	hakem28
8	6239	22	hakem15	hakem32	hakem28	55	8875	61	hakem38	hakem30	hakem35	102	3690	67	hakem24	hakem26	hakem33
9	6725	22	hakem24	hakem38	hakem20	56	6251	61	hakem29	hakem14	hakem32	103	5158	70	hakem30	hakem32	hakem14
10	8248	22	hakem32	hakem35	hakem14	57	1332	61	hakem12	hakem14	hakem29	104	6737	70	hakem15	hakem21	hakem25
11	3370	22	hakem26	hakem16	hakem25	58	2018	61	hakem35	hakem36	hakem24	105	9911	79	hakem18	hakem33	hakem22
12	1119	30	hakem22	hakem36	hakem23	59	6370	61	hakem14	hakem34	hakem19	106	4281	79	hakem13	hakem26	hakem16
13	7823	31	hakem17	hakem26	hakem16	60	6564	64	hakem24	hakem12	hakem25	107	6830	81	hakem23	hakem33	hakem22
14	9074	31	hakem31	hakem14	hakem15	61	9610	67	hakem27	hakem37	hakem32	108	2805	81	hakem16	hakem21	hakem20
15	9507	41	hakem34	hakem33	hakem36	62	8287	70	hakem35	hakem28	hakem23	109	7965	84	hakem25	hakem35	hakem36
16	3159	41	hakem20	hakem22	hakem21	63	7402	91	hakem28	hakem30	hakem26	110	8077	90	hakem20	hakem33	hakem24
17	5128	53	hakem22	hakem29	hakem28	64	3455	91	hakem38	hakem17	hakem22	111	8059	21	hakem37	hakem36	hakem19
18	7215	53	hakem15	hakem22	hakem35	65	1151	21	hakem18	hakem35	hakem28	112	6728	40	hakem13	hakem17	hakem32
19	4950	54	hakem29	hakem35	hakem38	66	1616	21	hakem19	hakem33	hakem12	113	4831	81	hakem17	hakem22	hakem19
20	7914	54	hakem29	hakem14	hakem25	67	4139	21	hakem29	hakem32	hakem24	114	8675	81	hakem25	hakem36	hakem27
21	1659	54	hakem15	hakem30	hakem27	68	7644	21	hakem34	hakem28	hakem25	115	6471	81	hakem35	hakem32	hakem25
22	2526	54	hakem20	hakem29	hakem30	69	4555	21	hakem19	hakem37	hakem23	116	5827	81	hakem21	hakem24	hakem15
23	1998	54	hakem26	hakem18	hakem34	70	9456	21	hakem33	hakem37	hakem31	117	2567	21	hakem25	hakem29	hakem13
24	6441	54	hakem17	hakem23	hakem29	71	7339	22	hakem11	hakem35	hakem36	118	3352	41	hakem35	hakem16	hakem20
25	5050	54	hakem35	hakem33	hakem19	72	3207	22	hakem19	hakem15	hakem28	119	4794	49	hakem32	hakem16	hakem28
26	4650	54	hakem34	hakem27	hakem15	73	9919	23	hakem34	hakem13	hakem32	120	3138	49	hakem35	hakem17	hakem12
27	8992	54	hakem35	hakem11	hakem20	74	8435	23	hakem37	hakem31	hakem22	121	1553	49	hakem16	hakem34	hakem23
28	3019	55	hakem20	hakem30	hakem19	75	7740	40	hakem30	hakem36	hakem19	122	9342	52	hakem25	hakem14	hakem30
29	2109	55	hakem30	hakem21	hakem19	76	3456	47	hakem33	hakem15	hakem28	123	7210	52	hakem38	hakem34	hakem23
30	1778	56	hakem38	hakem19	hakem12	77	8250	47	hakem11	hakem14	hakem15	124	4756	53	hakem12	hakem13	hakem33
31	6411	56	hakem34	hakem38	hakem11	78	5654	47	hakem30	hakem32	hakem37	125	6758	54	hakem25	hakem24	hakem29
32	1834	57	hakem33	hakem34	hakem11	79	6944	47	hakem15	hakem38	hakem14	126	4838	80	hakem32	hakem12	hakem17
33	7305	57	hakem32	hakem36	hakem17	80	5898	48	hakem18	hakem13	hakem26	127	8576	21	hakem28	hakem13	hakem25
34	7624	57	hakem13	hakem16	hakem17	81	9796	54	hakem20	hakem35	hakem26	128	8885	21	hakem12	hakem21	hakem31
35	1818	60	hakem27	hakem30	hakem24	82	9049	54	hakem30	hakem35	hakem29	129	3445	21	hakem11	hakem20	hakem17
36	5064	60	hakem24	hakem36	hakem29	83	2391	54	hakem16	hakem12	hakem27	130	1346	26	hakem17	hakem38	hakem11
37	6659	60	hakem37	hakem24	hakem22	84	7650	54	hakem22	hakem26	hakem30	131	8342	31	hakem19	hakem38	hakem24
38	4848	60	hakem18	hakem12	hakem15	85	3924	54	hakem32	hakem20	hakem37	132	8943	52	hakem33	hakem34	hakem11
39	7026	60	hakem34	hakem35	hakem12	86	8724	54	hakem21	hakem14	hakem15	133	5633	54	hakem18	hakem23	hakem27
40	2117	61	hakem11	hakem22	hakem21	87	7570	54	hakem18	hakem31	hakem26	134	3908	54	hakem20	hakem11	hakem34
41	3915	64	hakem36	hakem15	hakem22	88	3873	54	hakem37	hakem24	hakem38	135	5945	54	hakem26	hakem14	hakem38
42	2363	80	hakem32	hakem16	hakem27	89	1227	55	hakem19	hakem25	hakem14	136	9297	56	hakem33	hakem17	hakem18
43	9101	80	hakem23	hakem27	hakem20	90	6221	55	hakem21	hakem25	hakem13	137	5720	67	hakem29	hakem26	hakem27
44	3198	21	hakem22	hakem20	hakem28	91	3047	55	hakem16	hakem36	hakem13	138	1863	80	hakem21	hakem18	hakem29
45	6730	21	hakem19	hakem13	hakem34	92	4626	55	hakem14	hakem16	hakem33	139	1388	84	hakem16	hakem29	hakem34
46	4971	21	hakem17	hakem15	hakem12	93	1988	55	hakem23	hakem16	hakem38	140	6462	91	hakem37	hakem33	hakem20
47	9258	21	hakem36	hakem28	hakem38	94	8867	55	hakem35	hakem14	hakem24	141	2557	91	hakem38	hakem24	hakem22

142	2016	91	hakem12	hakem18	hakem13
143	5146	21	hakem20	hakem38	hakem16
144	8558	22	hakem25	hakem17	hakem33
145	6050	54	hakem34	hakem24	hakem33
146	6902	54	hakem24	hakem31	hakem16
147	3729	54	hakem27	hakem19	hakem38
148	5756	54	hakem20	hakem18	hakem13
149	7868	79	hakem37	hakem25	hakem19
150	9624	90	hakem26	hakem21	hakem33
151	1267	21	hakem11	hakem17	hakem13
152	1664	22	hakem36	hakem21	hakem30
153	6323	55	hakem19	hakem23	hakem12
154	5697	60	hakem21	hakem14	hakem34
155	4442	64	hakem27	hakem21	hakem31
156	9579	80	hakem31	hakem33	hakem17
157	7197	81	hakem13	hakem38	hakem31
158	5077	81	hakem18	hakem25	hakem21
159	3753	81	hakem24	hakem12	hakem21
160	5918	81	hakem13	hakem19	hakem35
161	6639	81	hakem29	hakem37	hakem14
162	3928	26	hakem22	hakem31	hakem30
163	6746	31	hakem24	hakem18	hakem17
164	8948	31	hakem36	hakem28	hakem23
165	4427	41	hakem16	hakem23	hakem18
166	9270	41	hakem31	hakem27	hakem21
167	6858	47	hakem36	hakem23	hakem20
168	5838	47	hakem25	hakem31	hakem17
169	9700	47	hakem28	hakem18	hakem34
170	6493	48	hakem30	hakem36	hakem35
171	6056	52	hakem22	hakem37	hakem21
172	2179	53	hakem30	hakem12	hakem14
173	4992	53	hakem13	hakem29	hakem31
174	2628	54	hakem21	hakem15	hakem12
175	4718	54	hakem36	hakem22	hakem35
176	8549	55	hakem12	hakem36	hakem18
177	1634	55	hakem36	hakem31	hakem22
178	2828	56	hakem36	hakem16	hakem19
179	3043	60	hakem14	hakem30	hakem19
180	7682	60	hakem15	hakem36	hakem38
181	7925	61	hakem27	hakem29	hakem26
182	5780	64	hakem28	hakem11	hakem37
183	9077	67	hakem21	hakem20	hakem33
184	4045	70	hakem23	hakem30	hakem37
185	8599	70	hakem29	hakem23	hakem28
186	3013	78	hakem32	hakem23	hakem12
187	4732	78	hakem16	hakem24	hakem26
188	5745	79	hakem15	hakem38	hakem22
189	1901	79	hakem19	hakem21	hakem30
190	5403	80	hakem31	hakem24	hakem32
191	2252	80	hakem29	hakem25	hakem26
192	5953	84	hakem23	hakem15	hakem24
193	7017	84	hakem33	hakem19	hakem25
194	6336	91	hakem38	hakem24	hakem31
195	7785	91	hakem17	hakem20	hakem18
196	8147	40	hakem28	hakem38	hakem21
197	5737	41	hakem11	hakem15	hakem25
198	1259	54	hakem11	hakem25	hakem22
199	5648	55	hakem23	hakem28	hakem27
200	5936	57	hakem30	hakem22	hakem19
201	6259	61	hakem24	hakem20	hakem37
202	6757	67	hakem25	hakem14	hakem15
203	6188	67	hakem17	hakem14	hakem18
204	8366	67	hakem34	hakem22	hakem20
205	1113	80	hakem14	hakem25	hakem17
206	6466	21	hakem26	hakem27	hakem38
207	8478	21	hakem14	hakem17	hakem24
208	6664	21	hakem21	hakem26	hakem14
209	1886	31	hakem30	hakem26	hakem25
210	7729	41	hakem36	hakem28	hakem16
211	9442	47	hakem18	hakem27	hakem21
212	5036	54	hakem21	hakem22	hakem16
213	3492	54	hakem13	hakem31	hakem29
214	2234	54	hakem22	hakem15	hakem33
215	6217	54	hakem13	hakem24	hakem25
216	6876	54	hakem16	hakem13	hakem36
217	7586	54	hakem21	hakem27	hakem31
218	5770	61	hakem25	hakem23	hakem15
219	8318	61	hakem26	hakem19	hakem31
220	4748	61	hakem29	hakem28	hakem18
221	5486	70	hakem12	hakem11	hakem16
222	9353	70	hakem15	hakem29	hakem37
223	8198	70	hakem17	hakem32	hakem11
224	4104	70	hakem19	hakem24	hakem34
225	7012	70	hakem27	hakem23	hakem12
226	6145	70	hakem20	hakem29	hakem30
227	6584	70	hakem23	hakem26	hakem11
228	8167	78	hakem27	hakem34	hakem20
229	7725	78	hakem32	hakem23	hakem11
230	9339	78	hakem28	hakem37	hakem35
231	4978	79	hakem11	hakem20	hakem32
232	8916	79	hakem27	hakem30	hakem11
233	7864	81	hakem31	hakem25	hakem38
234	5300	81	hakem22	hakem13	hakem27
235	1797	81	hakem34	hakem31	hakem29
236	2116	90	hakem35	hakem37	hakem36
237	5678	21	hakem16	hakem35	hakem37
238	3136	26	hakem22	hakem27	hakem36
239	2976	47	hakem11	hakem19	hakem31
240	5637	47	hakem23	hakem20	hakem36
241	8466	48	hakem33	hakem32	hakem34
242	2206	53	hakem14	hakem18	hakem23
243	5752	55	hakem29	hakem27	hakem12
244	9430	55	hakem18	hakem15	hakem36
245	3960	70	hakem32	hakem17	hakem37
246	3999	70	hakem13	hakem27	hakem28
247	1400	78	hakem31	hakem32	hakem16
248	4370	22	hakem38	hakem31	hakem13
249	3737	22	hakem31	hakem12	hakem35
250	1621	22	hakem25	hakem18	hakem28
251	7336	22	hakem37	hakem17	hakem36
252	1320	23	hakem17	hakem37	hakem13
253	1788	23	hakem31	hakem38	hakem17
254	9273	23	hakem28	hakem19	hakem17
255	9914	23	hakem26	hakem22	hakem37
256	2668	23	hakem30	hakem33	hakem23
257	8895	23	hakem38	hakem11	hakem32
258	7705	23	hakem15	hakem19	hakem32
259	6082	23	hakem19	hakem30	hakem16
260	8019	31	hakem18	hakem12	hakem33
261	7948	40	hakem26	hakem34	hakem36
262	5821	40	hakem20	hakem35	hakem13
263	2470	41	hakem34	hakem14	hakem11
264	4163	41	hakem14	hakem19	hakem13
265	4024	47	hakem32	hakem16	hakem33
266	3225	47	hakem33	hakem12	hakem18
267	7037	48	hakem16	hakem37	hakem14
268	9408	53	hakem12	hakem36	hakem27
269	7372	54	hakem26	hakem22	hakem20
270	5017	54	hakem37	hakem12	hakem38
271	2985	55	hakem21	hakem34	hakem29
272	8900	55	hakem37	hakem18	hakem30
273	3438	55	hakem12	hakem20	hakem35
274	5045	55	hakem27	hakem11	hakem35
275	3781	56	hakem38	hakem21	hakem26
276	3940	60	hakem23	hakem15	hakem28
277	9548	64	hakem24	hakem11	hakem37
278	9891	80	hakem34	hakem29	hakem25
279	3846	90	hakem14	hakem18	hakem17
280	2114	31	hakem15	hakem37	hakem11
281	2095	40	hakem26	hakem27	hakem38
282	9830	48	hakem14	hakem17	hakem24
283	5472	52	hakem21	hakem26	hakem14
284	7023	54	hakem30	hakem26	hakem25
285	8554	54	hakem36	hakem23	hakem16
286	6526	54	hakem18	hakem27	hakem21
287	4590	54	hakem21	hakem22	hakem16
288	3283	60	hakem13	hakem31	hakem29
289	2147	61	hakem22	hakem15	hakem33
290	8228	61	hakem13	hakem24	hakem28
291	3328	61	hakem16	hakem13	hakem36
292	9464	61	hakem35	hakem27	hakem31
293	9065	80	hakem25	hakem23	hakem15
294	9739	91	hakem26	hakem19	hakem31
295	1650	22	hakem38	hakem28	hakem18
296	9123	22	hakem12	hakem11	hakem16
297	9378	23	hakem15	hakem29	hakem37
298	9803	40	hakem17	hakem32	hakem11
299	8288	40	hakem19	hakem24	hakem34
300	8951	40	hakem27	hakem23	hakem12
301	7280	48	hakem20	hakem29	hakem30
302	1356	48	hakem23	hakem26	hakem11
303	5073	48	hakem27	hakem34	hakem20
304	6360	54	hakem32	hakem23	hakem11
305	7120	54	hakem36	hakem37	hakem35
306	6912	54	hakem11	hakem20	hakem32
307	2384	54	hakem32	hakem30	hakem11
308	8349	54	hakem31	hakem25	hakem38
309	9777	54	hakem22	hakem13	hakem27
310	1550	54	hakem38	hakem24	hakem22
311	5728	54	hakem12	hakem18	hakem13
312	1536	54	hakem20	hakem38	hakem16
313	5321	54	hakem25	hakem17	hakem33
314	5462	54	hakem34	hakem24	hakem33
315	4387	54	hakem24	hakem31	hakem16
316	5587	61	hakem27	hakem19	hakem38
317	9157	61	hakem20	hakem18	hakem13
318	4669	61	hakem37	hakem25	hakem19
319	1010	61	hakem26	hakem21	hakem33
320	1980	70	hakem11	hakem17	hakem13
321	6766	81	hakem36	hakem21	hakem30
322	2481	81	hakem19	hakem23	hakem12
323	6519	81	hakem35	hakem14	hakem34
324	3281	41	hakem27	hakem21	hakem31
325	7548	54	hakem31	hakem33	hakem17

EK2. Hakem ve Yazarlara ait Soru Sayıları

Hakem	Değ. soru say
hakem11	35
hakem12	35
hakem13	35
hakem14	34
hakem15	33
hakem16	35
hakem17	35
hakem18	34
hakem19	35
hakem20	35
hakem21	36
hakem22	34
hakem23	34
hakem24	36
hakem25	36
hakem26	35
hakem27	35
hakem28	31
hakem29	34
hakem30	34
hakem31	35
hakem32	35
hakem33	35
hakem34	35
hakem35	36
hakem36	37
hakem37	35
hakem38	36
Ortalama	34,821

Yazar	Yazdığı Soru say
Yazar 21	28
Yazar 22	16
Yazar 23	11
Yazar 26	3
Yazar 30	2
Yazar 31	9
Yazar 40	9
Yazar 41	10
Yazar 47	13
Yazar 48	9
Yazar 49	3
Yazar 52	5
Yazar 53	7
Yazar 54	54
Yazar 55	24
Yazar 56	5
Yazar 57	4
Yazar 60	10
Yazar 61	20
Yazar 64	5
Yazar 67	7
Yazar 70	15
Yazar 78	6
Yazar 79	7
Yazar 80	10
Yazar 81	17
Yazar 84	4
Yazar 90	4
Yazar 91	8
Ortalama	11,207

EK3. Kullanılan Yazılım

Usage

```
rm.sdt(dat, pid, rater, Qmatrix = NULL, theta.k = seq(-9, 9, len = 30),
    est.a.item = FALSE, est.c.rater = "n", est.d.rater = "n", est.mean=FALSE ,
    skillspace="normal" , tau.item.fixed = NULL , a.item.fixed = NULL ,
   d.min = 0.5, d.max = 100, d.start = 3, max.increment = 1, numdiff.parm = 0.00001,
   maxdevchange = 0.1, globconv = .001, maxiter = 1000, msteps = 4, mstepconv = 0.001)

## S3 method for class 'rm.sdt'
summary(object,...)

## S3 method for class 'rm.sdt'
plot(x, ask=TRUE, ...)

## S3 method for class 'rm.sdt'
anova(object,...)

## S3 method for class 'rm.sdt'
logLik(object,...)

## S3 method for class 'rm.sdt'
IRT.factor.scores(object, type="EAP", ...)

## S3 method for class 'rm.sdt'
IRT.irfprob(object,...)

## S3 method for class 'rm.sdt'
IRT.likelihood(object,...)

## S3 method for class 'rm.sdt'
IRT.posterior(object,...)

## S3 method for class 'rm.sdt'
IRT.modelfit(object,...)

## S3 method for class 'IRT.modelfit.rm.sdt'
summary(object,...)
```

Arguments

`dat`	Original data frame. Ratings on variables must be in rows, i.e. every row corresponds to a person-rater combination.
`pid`	Person identifier.
`rater`	Rater identifier.
`Qmatrix`	An optional Q-matrix. If this matrix is not provided, then by default the ordinary scoring of categories (from 0 to the maximum score of K) is used.
`theta.k`	A grid of theta values for the ability distribution.
`est.a.item`	Should item parameters a_i be estimated?
`est.c.rater`	Type of estimation for item-rater parameters c_{ir} in the signal detection model. Options are `'n'` (no estimation), `'e'` (set all parameters equal to each other), `'i'` (item wise estmation), `'r'` (rater wise estimation) and `'a'` (all parameters are estimated independently from each other).
`est.d.rater`	Type of estimation of d parameters. Options are the same as in `est.c.rater`.
`est.mean`	Optional logical indicating whether the mean of the trait distribution should be estimated.
`skillspace`	Specified θ distribution type. It can be `"normal"` or `"discrete"`. In the latter case, all probabilities of the distribution are separately estimated.
`tau.item.fixed`	Optional matrix with three columns specifying fixed τ parameters. The first two columns denote item and category indices, the third the fixed value. See Example 3.
`a.item.fixed`	Optional matrix with two columns specifying fixed a parameters. First column: Item index. Second column: Fixed a parameter.
`d.min`	Minimal d parameter to be estimated
`d.max`	Maximal d parameter to be estimated
`d.start`	Starting value of d parameters
`max.increment`	Maximum increment of item parameters during estimation
`numdiff.parm`	Numerical differentiation step width
`maxdevchange`	Maximum relative deviance change as a convergence criterion
`globconv`	Maximum parameter change
`maxiter`	Maximum number of iterations
`msteps`	Maximum number of iterations during an M step
`mstepconv`	Convergence criterion in an M step
`object`	Object of class `rm.sdt`
`x`	Object of class `rm.sdt`
`ask`	Optional logical indicating whether a new plot should be asked for.
`type`	Factor score estimation method. Up to now, only `type="EAP"` is supported.
`...`	Further arguments to be passed