T.C.
GAZİANTEP ÜNİVERSİTESİ
EĞİTİM BİLİMLERİ ENSTİTÜSÜ
EĞİTİM BİLİMLERİ ANA BİLİM DALI

STANDART VE SOLO TAKSONOMİSİNE DAYALI RUBRİKLER İLE PUANLANAN AÇIK UÇLU MATEMATİK SORULARINDA PUANLAYICI ETKİLERİNİN ÇOK YÜZEYLİ RASCH MODELİ İLE İNCELENMESİ

DOKTORA TEZİ

MUSTAFA İLHAN

GAZİANTEP
MART 2015

T.C.
GAZİANTEP ÜNİVERSİTESİ
EĞİTİM BİLİMLERİ ENSTİTÜSÜ
EĞİTİM BİLİMLERİ ANA BİLİM DALI

STANDART VE SOLO TAKSONOMİSİNE DAYALI RUBRİKLER İLE PUANLANAN AÇIK UÇLU MATEMATİK SORULARINDA PUANLAYICI ETKİLERİNİN ÇOK YÜZEYLİ RASCH MODELİ İLE İNCELENMESİ

DOKTORA TEZİ

MUSTAFA İLHAN

Tez Danışmanı: Doç. Dr. Bayram ÇETİN

GAZİANTEP
MART 2015

T.C.
GAZİANTEP ÜNİVERSİTESİ
EĞİTİM BİLİMLERİ ENSTİTÜSÜ
EĞİTİM BİLİMLERİ ANA BİLİM DALI

Standart ve SOLO Taksonomisine Dayalı Rubrikler ile Puanlanan Açık Uçlu Matematik Sorularında Puanlayıcı Etkilerinin Çok Yüzeyli Rasch Modeli ile İncelenmesi

MUSTAFA İLHAN

Tez Savunma Tarihi: 09.03.2015

Eğitim Bilimleri Enstitüsü Onayı

Doç.Dr. Mehmet Fatih ÖZMANTAR
Eğitim Bilimleri Enstitüsü Müdürü

Bu tezin Doktora tezi olarak gerekli şartları sağladığını onaylarım.

Prof.Dr. Zeynep HAMAMCI
Enstitü ABD Başkanı

Bu tez tarafımca okunmuş, kapsamı ve niteliği açısından bir Doktora tezi olarak kabul edilmiştir.

Doç.Dr. Bayram ÇETİN
Tez Danışmanı

Bu tez tarafımızca okunmuş, kapsam ve niteliği açısından bir Doktora tezi olarak kabul edilmiştir.

Jüri Üyeleri:

İmzası

Prof.Dr. Şener BÜYÜKÖZTÜRK (Jüri Başkanı) _____

Prof.Dr. Selahattin GELBAL _____

Doç.Dr. Bayram ÇETİN (Danışman) _____

Doç.Dr. Yılmaz SAĞLAM _____

Yrd. Doç.Dr. Yeşim ÖZER ÖZKAN _____

ÖZET

STANDART VE SOLO TAKSONOMİSİNE DAYALI RUBRİKLER İLE PUANLANAN AÇIK UÇLU MATEMATİK SORULARINDA PUANLAYICI ETKİLERİNİN ÇOK YÜZEYLİ RASCH MODELİ İLE İNCELENMESİ

İLHAN, Mustafa
Doktora Tezi, Eğitim Bilimleri ABD
Tez Danışmanı: Doç. Dr. Bayram ÇETİN
Mart 2015, 236 sayfa

Bu araştırmada, standart ve SOLO taksonomisine dayalı rubrikler ile puanlanan açık uçlu matematik sorularında puanlayıcı etkilerinin çok yüzeyli Rasch modeli ile incelenmesi amaçlanmıştır. Araştırmanın katılımcılarını, sekizinci sınıfa devam eden 104 ortaokul öğrencisi ile çalışmada puanlayıcı olarak görev alan yedi matematik öğretmeni oluşturmaktadır. Çalışmada veri toplama aracı olarak; araştırmacı tarafından geliştirilen ve açık uçlu sekiz sorudan oluşan matematik başarı testi, standart rubrikler, SOLO taksonomisine dayalı rubrikler, standart rubrikler ile ilgili düşünceler anketi (SRDA) ve SOLO taksonomisine dayalı rubrikler ile ilgili düşünceler anketi (STDRDA) kullanılmıştır. Araştırmada, açık uçlu matematik sorularına verilen yanıtların standart ve SOLO taksonomisine dayalı rubrikler kullanılarak puanlanmasıyla elde edilen veriler çok yüzeyli Rasch modeline göre analiz edilmiştir. Puanlayıcıların SRDA ile STDRDA'da yer alan kapalı uçlu maddelere verdikleri yanıtların analizinde aritmetik ortalama değerlerinden faydalanılmış; açık uçlu maddelere verilen yanıtlar ise betimsel analiz yaklaşımıyla çözümlenmiştir. Araştırmada; hem standart hem de SOLO taksonomisine dayalı rubriklere göre yapılan puanlamalarda yetenek düzeyi farklı olan öğrencilerin yüksek güvenirlikte birbirinden ayırt edilebildiği ve maddelerin güçlük düzeyleri arasında anlamlı fark olduğu belirlenmiştir. Standart rubrikler kullanılarak yapılan puanlamalarda; halo etkisi, tutarsızlık, *puanlayıcı*×*birey* ve *puanlayıcı*×*birey*×*madde* yanlılıklarının bulunmadığı saptanmıştır. Diğer taraftan; puanlamalarda merkeze yönelme etkisi ile *puanlayıcı*×*madde* yanlılığın söz konusu olduğu, katılık ve cömertlikleri yönüyle puanlayıcılar arasında anlamlı fark bulunduğu ve puanlayıcı güvenirliğinin düşük olduğu tespit edilmiştir. SOLO taksonomisine dayalı rubrikler kullanılarak yapılan puanlamalarda ise, puanlayıcı katılığı ve cömertliği, merkeze yönelme etkisi, halo etkisi, tutarsızlık ve yanlılık şeklinde sıralanan puanlayıcı etkilerinden hiç birine rastlanmamış; puanlayıcı güvenirliğinin yüksek olduğu ve puanlayıcıların benzer katılık/cömertlikte puanlamalar yaptığı sonucuna ulaşılmıştır.

Anahtar kelimeler: Açık uçlu matematik soruları, Puanlayıcı etkisi, Rubrik, SOLO taksonomisi, Çok yüzeyli Rasch modeli

ABSTRACT

THE IDENTIFICATION OF RATER EFFECTS ON OPEN-ENDED MATH QUESTIONS RATED THROUGH STANDARD RUBRICS AND RUBRICS BASED ON THE SOLO TAXONOMY IN REFERENCE TO THE MANY FACET RASCH MODEL

İLHAN, Mustafa
Ph.D. Dissertation, Department of Educational Sciences
Supervisor: Assoc. Prof. Dr. Bayram ÇETİN
March 2015, 236 pages

The purpose of the present study was to identify rater effects on open-ended math questions rated through standard rubrics and those based on the SOLO taxonomy in reference to the many facet Rasch model. The participants were 104 eight grade students and seven mathematics teachers serving as raters. The data collection instruments involved a mathematics achievement test comprised of eight open-ended questions composed by the researcher, standard rubrics, rubrics based on the SOLO taxonomy, the Survey of the Views on Standard Rubrics (SVSR), and the Survey of the Views on Rubrics based on the SOLO Taxonomy (SVRST). The data obtained from the rating of the responses to the open-ended math questions through standard rubrics and those rubrics based on the SOLO taxonomy were analyzed on the basis of the many facet Rasch model. On the other hand, the responses of the raters to the close-ended items included in the *SVSR* and *SVRST* were analyzed through arithmetic mean values whereas the responses to the open-ended items were analyzed descriptively. The findings suggested that both the standard rubrics and those based on the SOLO taxonomy could highly reliably distinguish among students of varying levels and that there was a significant difference between the items in terms of their difficulty. The rating through the standard rubrics did not suffer from the halo effect, inconsistency, *rater x individual* bias, or *rater x individual x item* bias. Nevertheless, the central tendency effect and *rater x item* bias were prevalent, there were significant differences among the raters in terms of severity and leniency, and the rater reliability was low. On the other hand, the rating through the rubrics based on the SOLO taxonomy were observed to be free from such rater effects as rater severity and leniency, the central tendency effect, the halo effect, inconsistency, and bias. Furthermore, the rater reliability was high, and the raters had similar levels of severity and leniency.

Key words: Open-ended math questions, Rater effects, Rubrics, SOLO taxonomy, Many facet Rasch model

ÖNSÖZ

Açık uçlu sorular puanlanışı bakımından subjektif testler arasında yer almaktadır. Subjektif testlerde öğrencinin yetenek düzeyine ilişkin geçerli ve güvenilir kestirimler elde edilebilmesi, ölçme sonuçlarını etkileyen puanlayıcı kaynaklı faktörlerin minimum düzeyde tutulmasına bağlıdır. Dolayısıyla, açık uçlu sorularda puanlayıcı etkilerinin minimum düzeyde tutulmasına yönelik farklı puanlama yöntemlerinin karşılaştırılması ve hangi puanlama yönteminin daha işlevsel olduğunun belirlenmesi bir gereklilik olarak karşımıza çıkmaktadır. Bu kapsamda araştırmada, standart ve SOLO taksonomisine dayalı rubrikler ile puanlanan açık uçlu sorularda puanlayıcı etkilerinin çok yüzeyli Rasch modeli ile incelenmesi amaçlanmıştır. Elde edilen bulgular ile getirilen önerilerin eğitimcilere ve araştırmacılara katkı sağlayacağını umduğum bu çalışma birçok değerli kişinin katkısı ile şekillenmiştir.

Öncelikle; araştırmanın her aşamasında görüş ve önerileriyle bana destek olan, bilimsel kişiliği ve tecrübelerinden çokça istifade ettiğim kıymetli danışman hocam Doç.Dr. Bayram ÇETİN'e teşekkür etmeyi bir borç bilirim. Doktora eğitimim boyunca aldığım derslerde sağladıkları katkılarla bilim insanı olma yolundaki naçizane çabamı destekleyen Prof.Dr. Hikmet YILDIRIM CELKAN'a, Doç.Dr. Erdal BAY'a, Doç.Dr. Birsen BAĞÇECİ'ye ve Doç.Dr. Servet DEMİR'e; yapıcı eleştirileri ile tezimin olgunlaşmasına katkıda bulunan Prof.Dr. Selahattin GELBAL'a, Prof.Dr. Şener BÜYÜKÖZTÜRK'e, Doç.Dr. Neşe GÜLER'e, Doç.Dr. Yılmaz SAĞLAM'a ve Yrd.Doç.Dr. Yeşim ÖZER ÖZKAN'a; yardım sever kişiliğiyle araştırma süresince desteğini esirgemeyen çalışma arkadaşım Arş.Gör. Melehat GEZER'e; sabır, hoşgörü ve şefkatleri ile daima yanımda olan aileme teşekkürlerimi sunarım. Son olarak, 2211 Yurt İçi Lisansüstü Burs Programı kapsamında doktora öğrenimim süresince maddi destek sağlayan TÜBİTAK'a teşekkür ederim.

Mustafa İLHAN

İÇİNDEKİLER

TABLOLAR LİSTESİ

ŞEKİLLER LİSTESİ

KISALTMALAR

MEB: : Milli Eğitim Bakanlığı

ÖSYM : Öğrenci Seçme ve Yerleştirme Merkezi

SRDA : Standart Rubrikler ile İlgili Düşünceler Anketi

STDRDA : SOLO Taksonomisine Dayalı Rubrikler ile İlgili Düşünceler Anketi

BİRİNCİ BÖLÜM

GİRİŞ

Bu bölümde araştırmanın problemi, amaç ve önemi, sınırlılıkları, sayıltıları ile araştırmaya ilişkin tanımlar yer almaktadır.

1.1. PROBLEM DURUMU

Öğrencilerin herhangi bir matematik konusunu gerçekte ne düzeyde anladığının belirlenmesi mümkün değildir. Öğrencinin konuyu ne düzeyde anladığına dair yapılan çıkarımlarda, bir göreve ilişkin ortaya koyduğu performans ya da kendisine yönetilen sorulara verdiği cevaplar etkili olmaktadır. Dolayısıyla, matematik performansının değerlendirilmesine yönelik çalışmaların altında yatan temel varsayım; öğrencinin geçerli ve güvenilir bir testte yer alan maddelere verdiği yanıtların, test ile ölçülmek istenen özellikleri ne düzeyde kazandığının geçerli bir göstergesi olacağıdır. Bu durum, öğrencilerin matematik başarılarının değerlendirilmesinde kullanılacak görev ve soruların seçimini, değerlendirme sürecinin oldukça kritik bir bileşeni haline getirmektedir (Romberg ve Wilson, 1992). Matematik değerlendirme sürecinde kullanılan yöntemler; *i)* matematiksel kavramları ve sistemleri anlama, *ii)* bu kavram ve sistemleri gerçek hayatta ve diğer öğrenme alanlarında kullanma, *iii)* tümevarım ve tümdengelim düşünce süreçlerini kullanarak çıkarımlar yapma, *iv)* matematiksel düşüncelerini açıklamak için matematiksel terminolojiyi doğru bir şekilde kullanma, *v)* problem çözme stratejileri geliştirme ve bunları günlük hayattaki problemlere uygulama gibi, matematik eğitimi kapsamında öğrencilere kazandırılması öngörülen becerilerin (MEB, 2009) öğrenciler tarafından ne düzeyde kazanıldığını ortaya çıkarabilecek nitelikte olmalıdır. Gerek sınıf içi değerlendirmelerde gerekse de geniş ölçekli sınavlarda sıklıkla kullanılan çoktan seçmeli testler, sıralanan ileri düzeydeki bilişsel davranışların ölçülmesinde yetersiz kalmaktadır (Kutlu, 2006). Dolayısıyla, matematik değerlendirme sürecinde çoktan seçmeli testlerin ötesinde yöntemlere

ihtiyaç duyulmaktadır. Bu ihtiyacı performans değerlendirmenin doğasında var olan özellikler karşılayabilmektedir (Güler, 2008). Performans değerlendirme, öğrencinin karmaşık problemleri çözmesini, problemi çözmek için kullandığı süreçleri göstermesini (McBee ve Barnes, 2009), cevabının gerekçelerini açıklamasını ve öğrendiklerini gerçek bir yaşam durumuna uygulamasını gerektirmektedir (Woodward, Monroe ve Baxter, 2001). Performans değerlendirmeye ilişkin bu özellikler, öğrencinin öğrenme sürecine daha aktif bir biçimde katılmasını, düşüncelerini daha özgür bir biçimde ifade etmesini, matematik bilgisini ve matematiksel düşünme becerisini kullanmasını ve öğrendiklerini birbiri ile ilişkilendirmesini sağlayarak (NAGB, 2002); karar verme, eleştirel düşünme, yaratıcı düşünme, analitik düşünme gibi üst düzey zihinsel becerilerinin gelişmesine yardımcı olmaktadır (Kind, 1999; Kutlu, Doğan ve Karakaya, 2010). Dolayısıyla, performans değerlendirmenin çoktan seçmeli testlere göre, günümüz toplumunda ihtiyaç duyulan karmaşık becerileri ve iletişim yeterliliklerini ölçmek için daha uygun olduğu söylenebilir (Palm, 2008).

Performans değerlendirmenin sıralanan avantajlarının yanı sıra bir takım sınırlılıkları bulunmaktadır. Bu sınırlılıklardan ilki; uygulanması ve puanlanmasının zaman alıcı olmasıdır. Kapsam geçerliğini sağlamanın zor oluşu performans değerlendirmeye ilişkin ikinci bir sınırlılıktır. Öğrencinin performans değerlendirme kapsamında sunulan görevleri yerine getirmesi ya da sorulan maddeleri yanıtlaması uzun süre gerektirmektedir. Bu durum, değerlendirmeye dâhil edilecek görevlerin sınırlı sayıda tutulmasına neden olmakta ve ölçme aracının kapsam geçerliğini olumsuz yönde etkilemektedir (Arias, 2010). Performans değerlendirmenin en önemli sınırlılığı ise, çoktan seçmeli testler gibi objektif bir biçimde puanlanamamasıdır (Romagnano, 2001). Öğrencilerin objektif olarak puanlanamayan herhangi bir testten aldığı puan, testi puanlayan kişiye göre farklılık gösterebilmektedir (Tekin, 2009). Alanyazında bu durumu örneklendiren çalışmalar bulunmaktadır. Söz gelimi; Özmantar, Bingölbali ve Akkoç (2008) tarafından yapılan araştırmada, 171 öğretmen açık uçlu bir matematik sorusuna verilen aynı öğrencinin cevabını puanlamış ve öğretmenlerin aynı cevaba sıfır ile 10 arasında yer alan geniş bir yelpazede oldukça farklı puanlar verdiği görülmüştür. Aynı cevaba yönelik olarak, öğretmenlerin %44'ü 10 üzerinden 10 tam puan verirken; %24'ü sıfır puan vermiştir. Bingölbali, Özmantar ve Akkoç (2008) tarafından yapılan bir başka araştırmada, öğrencilerin açık uçlu matematik sorularına verdikleri yanıtları

puanlarken, öğretmenlerin büyük bir çoğunluğunun kurala dayanan pratik çözümlere ayrıcalık tanıdığı; farklı çözüm yollarını göz ardı ettikleri sonucuna ulaşılmıştır. Performans değerlendirmede puanlayıcı farklılıklarına örnek teşkil edebilecek bir diğer çalışma Güler (2008) tarafından yapılmıştır. Güler (2008) tarafından yapılan araştırmada, öğrencilerin açık uçlu matematik sorularına verdikleri yanıtlar dört farklı puanlayıcı tarafından puanlanmıştır. Araştırma sonucunda, puanlamadaki katılık ve cömertlikleri açısından puanlayıcılar arasında anlamlı fark bulunduğu belirlenmiştir. Bu sonuçlar, açık uçlu sorularda puanlayıcılar ile ilgili faktörlerin öğrencinin performansını etkileyen değişkenlik kaynaklarından biri olduğunu göstermektedir.

Öğrencinin performansını etkileyen puanlayıcı kaynaklı faktörler *puanlayıcı etkisi* olarak adlandırılmaktadır (Farrokhi, Esfandiari ve Vaez Dalili, 2011). Puanlayıcı etkisi; puanlayıcıların katılığı ve cömertliği, halo etkisi, merkeze yönelme etkisi, ranj sınırlaması (Saal, Downey ve Lahey, 1980), tutarsızlık ve yanlılık olmak üzere altı başlıkta ele alınmaktadır (Myford ve Wolfe, 2004). Performans değerlendirmede, öğrencinin yetenek düzeyi hakkında doğru kestirimler elde edebilmek için sıralanan puanlayıcı etkilerinin kontrol altında tutulması gerekmektedir. Puanlayıcı etkilerini kontrol altına almak amacıyla, değerlendirme sürecinde birden fazla puanlayıcının görev alması ve puanlayıcı eğitimlerine başvurulması gibi farklı yöntemlerin işe koşulması önerilmektedir. Bu yöntemlerden biri rubrik (dereceli puanlama anahtarı) kullanımıdır (Dunbar, Brooks ve Miller, 2006). Rubrikler; puanlama işleminin, ne zaman ve kim tarafından yapıldığından bağımsız olarak gerçekleştirilmesini sağlar. Bu sayede puanlayıcı etkilerinin minimum düzeyde tutulmasına yardımcı olmaktadır (Moskal ve Leydens, 2000). Rubriklerin geliştirilmesinde yansıtıcı düşünme modeli, Bloom taksonomisi ya da SOLO taksonomisi gibi farklı modellerden yararlanılabilmektedir (Chan vd., 2002). Özellikle, SOLO taksonomisine dayalı rubrikler, açık uçlu soruları puanlamak için birçok farklı eğitim kademesinde ve birçok farklı derste sıklıkla kullanılmaktadır (Hattie ve Purdie, 1998).

SOLO taksonomisi gözlenebilir öğrenme çıktılarının yapısını açıklamak üzere ileri sürülmüştür. Yapı öncesi, tek yönlü yapı, çok yönlü yapı, ilişkisel yapı ve soyutlanmış yapı şeklinde beş düzeyden ulaşan bu taksonomisinin (Biggs, 2003) yaygın olarak kullanıldığı derslerden biri matematiktir (Collis ve Romberg, 1992; Lian ve Idris, 2006). SOLO taksonomisi; geometrik düşünme, cebirsel muhakeme ya

da gerçekçi matematik eğitimi gibi doğrudan matematik dersine yönelik olarak geliştirilmiş bir kuram olmasa da; matematik öğrenimi ve öğretimi ile matematik başarısının değerlendirilmesinde SOLO taksonomisinden yararlanılabilmektedir. Matematik dersi öğrenme çıktılarının SOLO taksonomisi ile değerlendirildiği araştırmalardan elde edilen bulgular; cebirsel düşünme, istatistiksel düşünme ve geometrik düşünme gibi matematik kapsamında yer alan farklı düşünme biçimlerinin SOLO taksonomisi ile ölçülebileceğini ortaya koymuştur. Jurdak (1991) tarafından yapılan çalışmada, SOLO taksonomisi ile Van Hiele geometrik düşünme düzeyleri arasında büyük ölçüde benzerlik olduğu belirlenmiş ve geometri dersi öğrenme çıktılarının değerlendirilmesinde SOLO taksonomisinden yararlanılabileceği sonucuna ulaşılmıştır. Mooney (2002) tarafından yapılan çalışmadan elde edilen bulgular, öğrencilerin istatistiksel düşünme süreçlerinin ölçülmesinde SOLO taksonomisinin uygun bir model olacağını göstermiştir. Lian ve Idris (2006), Lian, Yew ve Idris (2009), Lian, Meng, Yew ve Idris (2009) tarafından yapılan araştırmalarda ise öğrencilerin cebir problemlerini çözme becerilerinin değerlendirilmesinde SOLO taksonomisinden yararlanılabileceği tespit edilmiştir. Sonuç olarak; SOLO taksonomisinin cebir, istatistik ve geometri gibi matematiğin birçok farklı öğrenme alanında kullanıldığı söylenebilir. Dolayısıyla, SOLO taksonomisi matematik dersi öğrenme çıktılarının değerlendirilmesinde öncelikli olarak tercih edilebilecek bir model haline gelmektedir. Bununla birlikte, açık uçlu matematik sorularının puanlanmasında SOLO taksonomisine dayalı rubriklerin kullanımı konusunda nihai bir karar verilmeden önce bu rubriklerin puanlayıcı güvenirliği üzerindeki etkisinin ve puanlayıcı hatalarını kontrol altına alma konusundaki işlevselliğinin belirlenmesi gerekir.

Alanyazına bakıldığında, SOLO taksonomisine dayalı rubriklerin puanlayıcı güvenirliği üzerindeki etkisinin incelendiği sınırlı sayıda araştırma (Burnett, 1999; Chan, Tsui, Chan ve Hong, 2002; Hattie ve Purdie, 1998, Leung, 2000; Yazıcı, 2013) bulunduğu görülmektedir. Daha açık bir anlatımla; SOLO taksonomisine dayalı rubrik kullanılarak yapılan değerlendirmelerde puanlayıcı güvenirliğinin yüksek olacağına yönelik kuramsal bilgiler literatürde geniş bir yer tutsa da; bu rubriklerin puanlayıcı güvenirliğini nasıl etkilediğini ampirik olarak ortaya koyan araştırmaların sayısı oldukça sınırlıdır (Burnett, 1999; Chan vd., 2002; Hattie ve Purdie, 1998). Sözü geçen az sayıdaki araştırmada, SOLO taksonomisine dayalı rubriklerin puanlayıcı güvenirliği üzerindeki etkisine ilişkin çelişkili bulgular elde edilmiştir.

Örneğin; Hattie ve Purdie (1998), Burnett (1999) ve Chan vd. (2002), SOLO taksonomisine dayalı rubriklerin puanlayıcı güvenirliğini arttırdığını ifade ederken; Leung (2000) ve Chan, Hong ve Chan (2001), SOLO taksonomisine dayalı rubrik kullanarak yaptıkları değerlendirmelerde, puanlayıcılar arası güvenirliğin düşük olduğu sonucuna ulaşmıştır. SOLO taksonomisinin puanlayıcı güvenirliğine etkisini inceleyen araştırmalarda birbirinden farklı sonuçlar elde edilmesi, bu konuda yeni araştırmalara ihtiyaç olduğunu ortaya koymaktadır. Bu nedenle, değerlendirme sürecinde SOLO taksonomisine dayalı rubrik kullanımının, puanlayıcı güvenirliğini nasıl etkilediğini inceleyen ampirik araştırmaların literatüre kazandırılması bir ihtiyaç haline gelmektedir.

SOLO taksonomisine dayalı rubriklerin, puanlayıcı güvenirliğini nasıl etkilediği sorusuna isabet derecesi yüksek yanıtlar verebilmek için bu konuda literatüre kazandırılacak araştırmaların dikkatli bir biçimde planlanması gerekmektedir. Öncelikle, alanyazındaki çalışmalara bakıldığında, SOLO taksonomisine dayalı rubriklerin puanlayıcı güvenirliği üzerindeki etkisinin standart rubrikler ile karşılaştırılmadan incelendiği görülmektedir. Hangi rubrik türü kullanılırsa kullanılsın, rubriklerin puanlayıcılar arası farklılıkları azaltarak puanlayıcı güvenirliğini arttırması beklenmektedir (Airasian, 2005). Dolayısıyla, SOLO taksonomisine dayalı rubriklerin puanlayıcı hatalarını kontrol altına alma konusunda herhangi bir taksonomi temele alınmadan hazırlanan standart rubriklere kıyasla daha etkili olup olmadığının belirlenebilmesi için iki rubrik türüne göre yapılan puanlamaların karşılaştırmalı olarak incelenmesi gerekmektedir. Bu gereklilik dikkate alınmadan, SOLO taksonomisine dayalı rubriklerin puanlayıcı güvenirliği üzerindeki etkisinin puanlama sürecinde rubrik kullanımından mı; yoksa kullanılan rubrikte SOLO taksonomisinin temele alınmasından mı kaynaklandığı sorusunun yanıtlanması olanaklı değildir.

SOLO taksonomisine dayalı rubriklerin puanlayıcı güvenirliğini nasıl etkilediğini belirlemeye yönelik araştırmalarda, güvenirliğin hangi yöntem ile incelendiği oldukça önemli bir diğer konudur. Alanyazın incelendiğinde, SOLO taksonomisine dayalı rubriklerin puanlayıcı güvenirliğine etkisini araştıran çalışmalarda, basit uyum yüzdesi ve puanlayıcılar arası korelasyon katsayı gibi klasik test kuramına dayalı tekniklerin kullanıldığı görülmektedir. Buna bağlı olarak, söz konusu araştırmalar SOLO taksonomisine dayalı rubriklerin birey ve madde yüzeylerine ilişkin güvenirliği nasıl etkilediği sorusuna cevap olamamaktadır. Yine

aynı nedenle literatürdeki mevcut araştırmalar; SOLO taksonomisine dayalı rubriklerin puanlayıcı katılığı ve cömertliği, halo etkisi, merkeze yönelme etkisi, ranj sınırlaması, tutarsızlık ve yanlılık gibi puanlayıcı hataları üzerindeki etkisini ortaya koymada yetersiz kalmaktadır. Bu kapsamda, SOLO taksonomisine dayalı rubrik kullanımının sıralanan puanlayıcı hataları üzerindeki etkisinin incelenmesi gereklidir. SOLO taksonomisine dayalı rubrik kullanılarak puanlanan açık uçlu sorularda puanlayıcı etkilerinin çok yüzeyli Rasch modeli ile incelenmesiyle, değerlendirme işlemine karışan puanlayıcı etkileri belirlenebilir. Bu sayede, SOLO taksonomisine dayalı rubriklerin söz konusu puanlayıcı etkilerinin kontrol altına alınması amacına ne derece hizmet ettiği ortaya konulabilir.

1.2. ARAŞTIRMANIN AMACI

Bu araştırmada, standart ve SOLO taksonomisine dayalı rubrikler kullanılarak puanlanan açık uçlu matematik sorularında puanlayıcı etkilerinin çok yüzeyli Rasch modeli ile incelenmesi ve bu iki rubrik türüne ilişkin puanlayıcı görüşlerinin belirlenmesi amaçlanmıştır. Puanlayıcı katılığı ve cömertliği, merkeze yönelme etkisi, halo etkisi, tutarsızlık ve yanlılık araştırma kapsamında incelenen puanlayıcı etkileridir. Ranj sınırlaması; puanlayıcı katılığı, puanlayıcı cömertliği ya da merkeze yönelme şeklinde orta çıkan bir puanlayıcı etkisi olduğundan, bu etkinin ayrıca test edilmesine ihtiyaç duyulmamıştır. Çalışmanın amacı doğrultusunda aşağıdaki problemlere yanıt aranmıştır.

1) Standart rubriklere göre yapılan puanlamalarda; puanlayıcı, birey (öğrenci) ve madde yüzeyleri için hesaplanan güvenirlik değerleri ile uygunluk istatistikleri nasıldır?
2) Standart rubriklere göre yapılan puanlamalarda; *i)* puanlayıcı katılığı ve cömertliği, *ii)* merkeze yönelme etkisi, *iii)* halo etkisi, *iv)* tutarsızlık ve *v)* yanlılık bulunmakta mıdır?
3) SOLO taksonomisine dayalı rubrikler kullanılarak yapılan puanlamalarda; puanlayıcı, birey (öğrenci) ve madde yüzeyleri için hesaplanan güvenirlik değerleri ile uygunluk istatistikleri nasıldır?
4) SOLO taksonomisine dayalı rubrikler kullanılarak yapılan puanlamalarda; *i)* puanlayıcı katılığı ve cömertliği, *ii)* merkeze yönelme etkisi, *iii)* halo etkisi, *iv)* tutarsızlık ve *v)* yanlılık bulunmakta mıdır?

5) Standart ve SOLO taksonomisine dayalı rubrik kullanılarak yapılan puanlamalardan elde edilen yetenek kestirimleri ile öğrencilerin matematik karne notları ve 2013-2014 Öğretim Yılı matematik dersi ortak sınavındaki doğru sayıları arasındaki korelasyonlar nasıldır?
6) Standart ve SOLO taksonomisine dayalı rubrik kullanılarak yapılan puanlamalardan elde edilen yetenek kestirimleri arasında anlamlı fark var mıdır?
7) Puanlayıcıların, standart ve SOLO taksonomisine dayalı rubrikler hakkındaki görüşleri nasıldır?

1.3. ARAŞTIRMANIN ÖNEMİ

Türkiye'de 2005 yılından itibaren ilköğretimde, 2007 yılından itibaren ortaöğretimde uygulanmaya başlanan matematik dersi öğretim programlarında; matematiği günlük hayata uygulama, problem çözme ve akıl yürütme becerilerinin öğrencilere kazandırılması matematik eğitiminin genel amaçları arasında sayılmıştır. Bu amaçların öğrenciler tarafından ne düzeyde kazanıldığının tespit edilebilmesi için, matematik değerlendirme sürecinde tek bir doğru cevabı olan çoktan seçmeli testler yerine; öğrencilerin değişik çözüm yollarını kullanmasına ve farklı sonuçlara ulaşmasına olanak tanıyan açık uçlu sorulardan yararlanılması önerilmektedir (MEB; 2009; MEB; 2011). Ayrıca, MEB ilerleyen yıllarda ulusal düzeydeki sınavlarda açık uçlu sorulara yer vermeyi planlamaktadır. MEB tarafından temel eğitimden ortaöğretime geçiş sistemi ile ilgili olarak yapılan açıklamada, FATİH projesi tüm bileşenleriyle birlikte uygulamaya geçtiğinde, her öğrencinin elinde tablet bilgisayar olacağı ve öğrencilerin elindeki tablet bilgisayarların geniş ölçekli sınavlarda açık uçlu soruların sorulmasına imkân tanıyacağı ifade edilmiştir (MEB, 2013). Aynı şekilde, Öğrenci Seçme ve Yerleştirme Merkezi (ÖSYM) 2013 yılı itibariyle, Açık Uçlu Sorularla Sınav Projesi'ni başlatmıştır. Bu proje kapsamında, ilerleyen yıllarda aday sayısının az olduğu sınavlar başta olmak üzere merkezi sınavlarda çoktan seçmeli soruların yanı sıra açık uçlu sorulara da yer verilmesi düşünülmektedir. Ancak açık uçlu sorular, çoktan seçmeli testler gibi objektif bir biçimde puanlanamamaktadır (Tuckman, 1991). Öğrencinin herhangi bir açık uçlu sorudan alacağı puan, puanlamayı yapan kişiye göre farklılık gösterebilmektedir. Öğrencinin başarısı hakkında yapılan değerlendirmelerin güvenilir olabilmesi için öğrencinin yetenek düzeyi dışında, performans puanlarında değişkenliğe neden olabilecek bütün

faktörlerinin minimum düzeyde tutulması gerekmektedir (Cardinet, Johnson ve Pini, 2010). Bu bağlamda; açık uçlu soruların daha objektif bir biçimde puanlamasına olanak tanıyacak alternatif yöntemlerin etkililiğinin incelenmesi, öğrenci başarısının güvenilir bir biçimde kestirilebilmesi için oldukça önemlidir. Açık uçlu soruların puanlamasında kullanılabilecek iki alternatif yöntem olan standart rubrikler ile SOLO taksonomisine dayalı rubriklerin karşılaştırılması ve puanlayıcı hatalarını kontrol altına alma konusunda daha etkili olduğu belirlenen yöntemin uygulamada işe koşulması, öğrenci başarısının daha güvenilir bir biçimde belirlenmesine yardımcı olacaktır. Bu yönüyle araştırmanın uygulamaya yönelik önemli doğurgularının olacağı öngörülmektedir.

Uygulamaya yönelik doğurgularının yanı sıra, araştırmanın ilgili literatüre de önemli katkılar sağlayacağı tahmin edilmektedir. Alanyazında SOLO taksonomisine dayalı rubriklerin halo etkisi, ranj sınırlaması, merkezi yönelme etkisi, puanlayıcı katılığı ve cömertliği gibi puanlayıcı etkileri açısından analiz edildiği bir çalışmaya rastlanmadığı dikkate alındığında; araştırmanın ölçme değerlendirme alanına bilimsel yenilik getireceği söylenebilir. Ayrıca araştırma sayesinde, puanlayıcı etkilerinin tespitinde çok yüzeyli Rasch modelinin kullanımına örnek teşkil edecek Türkçe bir kaynak literatüre kazandırılmış olacaktır. Türkçe literatürde puanlayıcı hataları ve puanlayıcı eğitimleri ile ilgili kuramlar bilgilerin yer aldığı çalışmalar (İlhan ve Çetin 2014a; İlhan ve Çetin, 2014b) bulunsa da; puanlayıcı etkilerinin nasıl tespit edilebileceğini gösteren bir çalışmaya rastlanmıştır. Türkçe literatürdeki bu boşluğu dolduracak olması araştırmayı önemli kılan bir diğer özelliktir. Son olarak; bu çalışma açık uçlu matematik soruları üzerinden yürütüldüğünden araştırmada ulaşılacak bulgular matematik eğitimi ile ilgili literatüre de önemli katkılar sunacaktır. Bununla birlikte, SOLO taksonomisi içerikten bağımsız bir model olduğundan (Kanuka, 2011); araştırma sonuçlarının dolaylı olarak farklı disiplinlerdeki ölçme-değerlendirme çalışmalarına da kılavuzluk etmesi beklenmektedir.

1.4. SAYILTILAR

Puanlayıcılar, öğrencilerin açık uçlu matematik sorularına verdikleri yanıtları ciddiyetle ve birbirlerinden bağımsız olarak puanlamışlardır.

1.5. SINIRLILIKLAR

Araştırmaya ilişkin sınırlılıklar kavramsal sınırlılıklar ve yöntemsel sınırlılıklar olmak üzere iki başlıkta sunulmuştur.

1.5.1. Kavramsal Sınırlılıklar

Bu araştırmada incelenen puanlayıcı etkileri; puanlayıcı katılığı ve cömertliği, merkeze yönelme etkisi, halo etkisi, tutarsızlık ve yanlılık ile sınırlı tutulmuştur.

1.5.2. Yöntemsel Sınırlılıklar

Araştırmada, standart ve SOLO taksonomisine dayalı rubriklerin puanlayıcı etkileri açısından karşılaştırılmasına yönelik veriler, 104 öğrencinin açık uçlu sekiz matematik sorusuna verdikleri yanıtın yedi puanlayıcı tarafından puanlanması sonucu elde edilmiştir. Dolayısıyla, çok yüzeyli Rasch analizleri sadece yukarıda verilen sayıda öğrenci, madde ve puanlayıcıdan elde edilen veriler ile sınırlı kalmıştır.

1.6. TANIMLAR

Standart Rubrik: Araştırma kapsamında, herhangi bir taksonomi temele alınmadan hazırlanan rubrikler standart rubrik olarak isimlendirilmiştir.

SOLO Taksonomisine Dayalı Rubrik: SOLO taksonomisinin yapı öncesi, tek yönlü yapı, çok yönlü yapı, ilişkisel yapı ve soyutlanmış yapı düzeyleri referans alınarak oluşturulan rubrikler SOLO taksonomisine dayalı rubrik olarak adlandırılmaktadır.

İKİNCİ BÖLÜM

KAYNAK ÖZETLERİ

Bu bölümde, ilk olarak araştırma problemi ile ilgili kuramsal çerçeve sunulmuştur. Daha sonra, konu ile ilgili yurt içi ve yurt dışında yapılan araştırmalara yer verilmiştir.

2.1. KURAMSAL ÇERÇEVE

2.1.1. Matematik Başarısının Değerlendirilmesi

Değerlendirme, öğrencinin performansı hakkında farklı ölçme araçları ile toplanan verilere dayalı olarak yapılan bir yorumlama işlemidir (Brown ve Hirschfeld, 2008). Değerlendirme çalışmalarıyla öğrenme sürecinin etkililiğinin gözden geçirilmesi (Chamoso ve Caceres, 2009), öğrencilerin gelişimlerinin takip edilmesi ve öğrenme ile ilgili ihtiyaçlarının belirlenmesi amaçlanmaktadır (Altun, 2005). Değerlendirme çalışmalarının tüm bilim dalları için geçerli olan bir takım ortak amaçları bulunmasına rağmen, değerlendirme ile ilgili standartlar disiplinlere göre farklılık göstermektedir (Alkan, 1999). Matematik dersi söz konusu olduğunda, değerlendirme standartlarının belirlenmesinde, büyük ölçüde Ulusal Matematik Öğretmenleri Konseyi'nin (The National Council of Teachers of Mathematics-NCTM) kararları etkili olmaktadır. NCTM, 1989 yılında yayınladığı raporda matematik dersine ilişkin iddialı reformlar önermiştir. Matematik dersinin amaçları, matematik öğretiminde nasıl bir yol takip edilmesi gerektiği ve matematik dersine ilişkin değerlendirme standartları bu raporun genel çerçevesini oluşturmaktadır. Bu rapora göre, öğrencilerin akıl yürütme, problem çözme ve matematiksel iletişim kurma becerilerinin geliştirilmesi matematik eğitiminin temel amaçlarını oluşturmaktadır. Matematik dersinin hedeflerindeki bu değişiklikler öğrencilerin nasıl değerlendirilmeleri gerektiğine ilişkin değişiklikleri beraberinde getirmiştir. Dolayısıyla, 1989 tarihli belgenin bir bölümünü matematik dersinin değerlendirilmesine yönelik değişiklikler oluşturmaktadır.

Matematik dersinin değerlendirilmesine ilişkin esas değişiklikler ise, NCTM'nin 1995 yılında yayınladığı ve tamamıyla değerlendirme ile ilgili konulara ayrılan "Okul Matematiği için Değerlendirme Standartları" adlı rapor ile gerçekleşmiştir. Bu raporda, öğrencilerin matematik öğrenmelerine ilişkin geçerli kanıtlar elde edilebilmesi için performans görevleri, projeler ve portfolyolar gibi çoklu değerlendirme kaynaklarından yararlanılması önerilmiştir. Matematik değerlendirme sürecinde, çoktan seçmeli testler yerine öğrencilerin çözüme nasıl ulaştıklarını gösterebilecekleri ve çözümlerini gerekçelendirebilecekleri açık uçlu sorulardan yararlanılması, üzerinde önemle durulan bir diğer konudur (Archbald ve Grant, 2000). Çünkü çoktan seçmeli testlerin, öğrencilerin matematik başarıları hakkında geçerli değerlendirmeler yapmaya engel teşkil eden bir takım sınırlılıkları bulunmaktadır (Woodward, Monroe ve Baxter, 2001).

İlk olarak çoktan seçmeli testlerde, madde ile ölçülen özelliğe sahip olmayan ya da kısmen sahip olan öğrenciler tahmin ile doğru cevaba ulaşabilmektedir. Şans başarısı olarak adlandırılan bu durum testin güvenirliğini ve geçerliğini olumsuz yönde etkilemektedir (Turgut ve Baykul, 2012). Çoktan seçmeli sınavlara ilişkin ikinci bir sınırlılık, bu sınavlarda öğrencilerin kopya çekme olasılığının diğer pek çok sınav türüne göre daha yüksek olmasıdır (Alharby, 2006). Çoktan seçmeli testlere ilişkin üçüncü bir sınırlılık, bu sınavlardaki maddeler için yapılan doğru-yanlış şeklindeki değerlendirmelerin kısmi puanlamaya uygun olmamasıdır (Bağcan Büyükturan ve Çıkrıkçı Demirtaşlı, 2013). Çoktan seçmeli testlere ilişkin bu sınırlılıkların aşılması imkânsız olmadığından bu dezavantajlar, çoktan seçmeli testlerin ikinci dereceden sınırlılıkları olarak nitelendirilmektedir (Alharby, 2006). Örneğin, şans başarısını kontrol altında tutmak için düzeltme formülü uygulanabilmekte (Doğan, 2009; Gültekin, 2012) ya da madde güçlük ve ayıcılık parametreleriyle birlikte şans parametresini de içeren üç parametreli madde tepki kuramı modelinden yararlanılabilmektedir (Osterlind, 2002). Sınav öncesinde ve sınav sırasında gerekli önlemler alınarak öğrencilerin kopya çekme olasılıkları da kontrol altında tutulabilmektedir. Çoktan seçmeli testlerde geleneksel yöntemle yapılan puanlamalar yerine eleme yöntemiyle yapılan puanlamalar kullanılarak, bu testler kısmi puanlamalar için uygun hale getirilebilmektedir. Eleme yöntemiyle yapılan puanlamada, öğrenciden yanlış olduğunu düşündüğü seçenekleri elemesi istenmekte ve elediği seçenek sayısına göre bir puanlama yapılmaktadır (Çetin, 2005). Görüldüğü gibi, çoktan seçmeli testlere ilişkin bu sınırlılıklar, farklı puanlama

yöntemleri kullanılarak ve gerekli önlemler alınarak kontrol altında tutulabilmektedir. Diğer taraftan, çoktan seçmeli testlere ilişkin bir takım sınırlılıkların üstesinden gelmek çok daha zor veya imkânsızdır. Bu sınırlılıklar, çoktan seçmeli testlerin, öğrencilerin gerçekte ne kadar iyi öğrendiğini yansıtmada yetersiz kalan yapısından kaynaklanmaktadır (Romagnano, 2001).

Çoktan seçmeli testlerde öğrenci cevabı kendisi oluşturmamaktadır. Öğrencinin sunulan seçeneklerden uygun olanı belirleyerek doğru cevaba ulaşması esastır (Birenbaum ve Feldman, 1998; Simkin ve Kuechler, 2005). Ancak, gerçek hayatta bireylerin matematik bilgilerini kullanmalarını gerektiren koşullar, sunulan seçeneklerden herhangi birini seçmesini gerektiren durumlardan çok daha farklıdır. Örneğin, bir tüccarın gelir ve giderlerini hesaplaması veya bir banka yetkilisinin müşterinin sunduğu ipoteğin konut kredisi için yeterli olup olmadığını belirlemesi gibi matematik bilgisini kullanmayı gerektiren birçok durumda, birey sunulan seçeneklerden birini seçmek yerine cevabı kendisi yapılandırmaktadır (Stecher, 2010). Dolayısıyla, çoktan seçmeli testler, günlük hayatta karşılaşacağı matematik problemlerine çözüm üretebilecek donanıma sahip bireyler yetiştirmede yetersiz kalmaktadır. Çoktan seçmeli testlerde, yalnızca bireyin ne bildiği ya da bir işlemi nasıl sonuçlandırdığıyla ilgilenilmektedir. Diğer bir deyişle; çoktan seçmeli testlerde, açıklayıcı ya da işlemsel bilgiye odaklanılmaktadır. Öğrencinin konu ile ilgili sistematik ve stratejik bilgisini belirlemede ise çoktan seçmeli testler yetersiz kalmaktadır. Daha açık bir anlatımla; öğrencinin bildiklerini ne zaman, nasıl, nerede ve niçin uygulayacağı ile yeterlilikleri çoktan seçmeli testler kullanılarak ölçülememektedir (Stecher, 2010). Çoktan seçmeli testlerin diğer bir dezavantajı da, öğrenmeye olan olumsuz etkisidir. Çoktan seçmeli testlerde, öğrencinin testteki sorulardan kaç tanesini doğru yanıtladığı başarısının tek göstergesi olarak kabul edilmektedir (Romberg ve Wilson, 1992). Bu durum öğretmenlerin, ailelerin ve öğrencilerin yalnızca bu testlerdeki başarıya odaklanmasına ve test ile ölçülen özelliklerin gerçekten kazanılıp kazanılmadığını göz ardı etmesine neden olmaktadır. Yine bu durum, çoktan seçmeli testler ile ölçülmesi mümkün olmayan daha önemli becerilerin öğrencilere kazandırılmasına engel teşkil etmektedir (Alharby, 2006). Çoktan seçmeli testlere ilişkin bu sınırlılıkların üstesinden gelebilmek için performans değerlendirmeden yararlanılması önerilmektedir (Kantrov, 2000).

2.1.2. Performans Değerlendirme

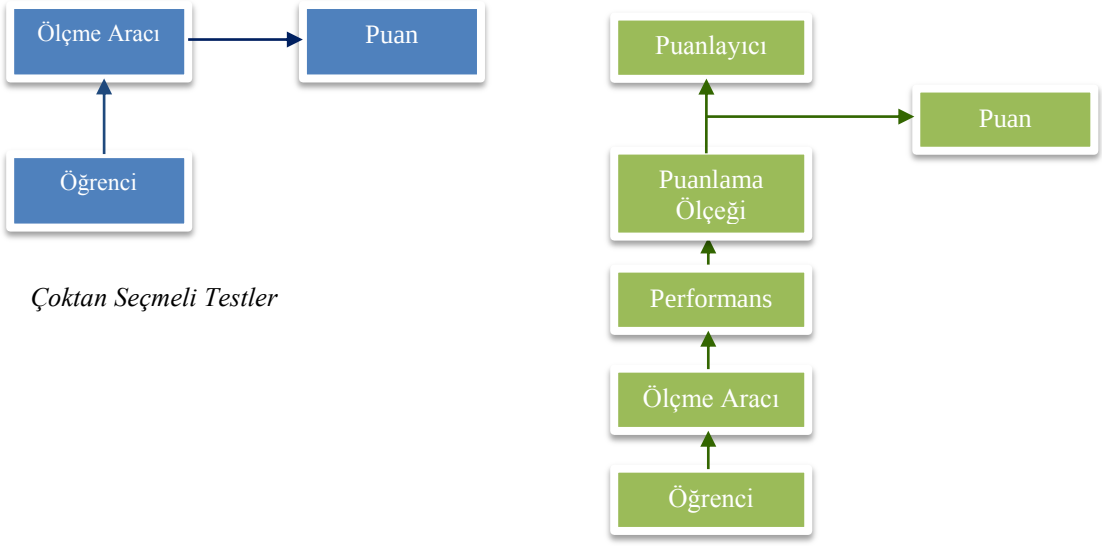
Performans değerlendirme farklı araştırmacılar tarafından değişik biçimlerde tanımlanmıştır. Araştırmacıların performans değerlendirmeye birbirinden oldukça farklı anlamlar yüklemesi, performans değerlendirmenin sınırlarını çizmeyi güçleştirmektedir (Palm, 2008). Stecher (2010) performans değerlendirmenin sınırlarını net bir biçimde çizmek için performans değerlendirmeyi tanımlamak yerine; performans değerlendirmenin ne olmadığına odaklanmayı önermiştir. Performans değerlendirme; çoktan seçmeli bir test, doğru yanlış testi ya da eşleştirme testi değildir (Danielson ve Hansen, 1999; Stecher, 2010). Performans değerlendirmede, öğrencinin sunulan seçeneklerden herhangi birini seçmesi yerine; yanıtı kendisinin oluşturması gerekmektedir (Zhu, 2009). Dolayısıyla, öğrencinin bilgiyi hafızadan geri getirmesi ile ilgilenen çoktan seçmeli testlerin aksine; performans değerlendirme bilginin öğrenci tarafından yapılandırılması ile ilgilenmektedir (Moore, 2009). Bu yönüyle performans değerlendirme öğrencinin güçlü ve zayıf yönlerini belirleme ve öğrendiklerine ilişkin daha ayrıntılı bilgi sahibi olma imkânı sunmaktadır (Khattri, Kane ve Reeve, 1995). Performans değerlendirmede, öğrenciden çözüme ulaşmasının yanı sıra çözüme nasıl ulaştığını açıklaması ve cevabını gerekçelendirmesi de istenmektedir. Performans değerlendirmenin bu özelliği öğrencilerin matematiksel kavramlara ilişkin daha derin öğrenmeler gerçekleştirmesini sağlamakta (Woodward, Monroe ve Baxter, 2001) ve eleştirel düşünme becerilerinin gelişmesine yardımcı olmaktadır (Burns, 1995). Performans değerlendirmenin bir diğer avantajı da, öğrencinin üst düzey bilişsel süreçlerine odaklanmasıdır (Brualdi, 1998; Johnson, Penny ve Gordon, 2009). Performans değerlendirme, söz konusu olumlu özellikleriyle birlikte, sistematik ve tesadüfi (random) hatalardan etkilenmeye açık doğası (Cronbach, 1990; Yue, 2011) nedeniyle ölçümlerin geçerliğine ilişkin tehditleri beraberinde getirmektedir (Abu Kassim, 2007).

Performans değerlendirme, ölçülen özelliğin dışında; öğrenciden gerçekleştirmesi beklenen görevin doğası, öğrencinin sunulan probleme ya da göreve olan ilgisi ve bu konudaki alt yapısı gibi birçok farklı kaynaktan ve bu kaynakların etkileşimiyle oluşan çeşitli faktörlerden etkilenmektedir. Sıralanan dışsal kaynaklardan dolayı puanlamada oluşan değişkenlikler ölçme hatası olarak ifade edilmektedir (Sudweeks, Reeveb ve Bradshawc, 2005). Öğrencinin performansına ilişkin çıkarımları etkileyen ve ölçme hatasına neden olan dışsal değişkenlikler

değerlendirme sonuçlarının güvenilirliğini zedelemektedir (Huang, 2009). Ölçülen yapı ile ilgili olmayan varyansa yol açan dışsal kaynaklardan biri de puanlayıcılar ile ilgili faktörlerdir (Brennan, Gao ve Colton, 1995; Congdon ve McQueen, 2010). Puanlayıcının ana dili (Hamp-Lyons ve Zhang, 2001; Kobayashi, 1992), puanlama tecrübesi (Cumming, 1990), cinsiyeti ve daha önce aldığı puanlama eğitimleri gibi faktörler puanlayıcıların değerlendirme sürecindeki davranışlarını etkileyebilmektedir (Barkaoui, 2007). Öğrencilerin performans puanlarını etkileyen puanlayıcı kaynaklı faktörler *puanlayıcı etkisi* olarak adlandırılmaktadır (Farrokhi, Esfandiari ve Vaez Dalili, 2011; Schaefer, 2008).

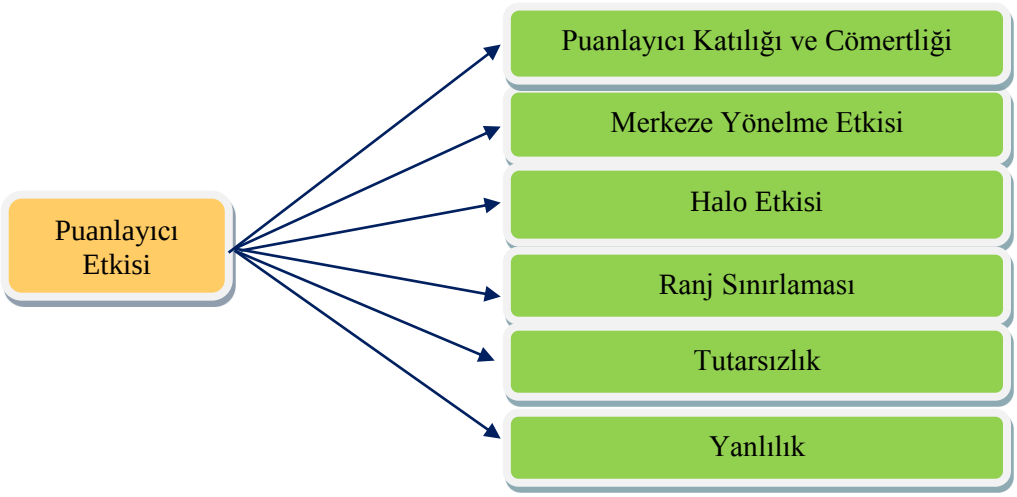
2.1.3. Puanlayıcı Etkisi

Çoktan seçmeli testler gibi objektif olarak puanlanabilen testler yüksek puanlama güvenirliğine sahiptir. Çünkü öğrencilerin bu tür bir testten aldıkları puanlar testi puanlayan kişiye göre değişmemektedir. Oysa açık uçlu sorular gibi, objektif olarak puanlanamayan ölçme araçları için aynı şeyi söylemek güçtür. Objektif olarak puanlama yapmanın mümkün olmadığı değerlendirmelerde öğrenciye verilen puan, puanlamayı yapan kişiye göre farklılık gösterebilmektedir (Tekin, 2009). Dolayısıyla, performans değerlendirmede, puanlayıcılar ölçme aracından elde edilen puanlar için bir değişkenlik kaynağı olmaktadır. Bu durum, performans değerlendirmede puanlayıcı etkilerini göz ardı edilemeyecek bir problem haline getirmektedir (Eckes, 2005; Shrout ve Fleiss, 1979).



Şekil 2.1. Çoktan seçmeli testler ile performans değerlendirmede puanlama süreci (Ramineni, 2008)

Puanlayıcı etkisi, öğrencinin performans puanlarına ölçülen yapı ile ilgisi olmayan varyansın karışmasına neden olan, puanlayıcılara ait istenmeyen sistematik davranışlardır (Bachman, 2004; Eckes, 2005; Hoyt, 2000; Scullen, Mount ve Goff, 2000). Puanlayıcı etkisi; puanlayıcıların katılığı ve cömertliği, halo etkisi, merkeze yönelme etkisi, ranj sınırlaması (Saal, Downey ve Lahey, 1980), tutarsızlık ve yanlılık olmak üzere altı kategoride incelenmektedir (Myford ve Wolfe, 2004).



Şekil 2.2. Puanlayıcı etkisinin türleri

2.1.3.1. Puanlayıcı Katılığı ve Cömertliği

Puanlayıcı katılığı ve cömertliği, puanlayıcılardan birinin sürekli olarak diğer puanlayıcılardan veya puanlama ölçütlerinden daha düşük ya da daha yüksek puan verme eğiliminde olmasıdır (Jackson, Schuler ve Werner, 2009). Saal, Downey ve Lahey (1980), puanlayıcı katılığını, puanlayıcının puanlama ölçeğinin orta noktasının altında; puanlayıcı cömertliğini ise puanlayıcının puanlama ölçeğinin orta noktasının üzerinde bir puan verme eğiliminde olması şeklinde tanımlamıştır. Puanlayıcının katılığı ve cömertliği; performansı, katı bir puanlayıcı tarafından değerlendirilen bir öğrencinin, ölçülen özellik açısından kendisinden daha az yetenekli olan ancak daha cömert bir puanlayıcı tarafından değerlendirilen bir öğrenciye göre daha düşük bir puan almasına neden olabilmektedir (Wiseman, 2012a). Bu durumda, iki öğrencinin puanlarındaki varyans sadece öğrencilerin yetenek düzeylerini yansıtmamakta; puanlayıcıların katılık ve cömertliklerindeki farklılıkları da içermektedir. Puanlayıcı katılığı özellikle, puanları kesme noktasında olan öğrenciler için telafisi zor alan ciddi sonuçlara neden olabilmektedir. Örneğin; bir son sınıf öğrencisinin performansının katı bir puanlayıcı tarafından değerlendirilmesi öğrencinin dönem

uzatması ya da yıl kaybetmesi anlamına gelebilir (McNamara, 1996). Puanlayıcı katılığı ve cömertliği, Cronbach (1990) tarafından puanlama sürecine karışan en önemli puanlayıcı etkisi olarak nitelendirilmiştir. Tablo 2.1'de (Wolfe, 2014), puanlayıcı katılığı ve cömertliğine örnek teşkil eden puanlamalar sunulmuştur.

Tablo 2.1. Puanlayıcı katılığı ve cömertliğine örnek teşkil eden puanlamalar

Gerçek Puanlar	0	1	1	2	2	2	3	3	3	3	3	4	4	4	5	5	6
Puanlayıcı Cömertliği	2	2	2	3	4	4	3	5	5	4	6	6	5	5	6	6	6
Puanlayıcı Katılığı	0	0	0	1	2	1	1	3	2	2	3	3	2	2	4	4	5

Tablo 2.1'in ilk satırında gerçek puanlar yer almaktadır. Tablonun ikinci satırında sunulan puanlamalarda, puanlayıcı cömertliği; üçüncü satırında sunulan puanlamalarda ise puanlayıcı katılığı söz konusudur. Tabloda gösterilen puanlamalardan da anlaşılacağı üzere, cömert puanlayıcılar tarafından yapılan puanlamaların ortalaması, gerçek puanlara göre daha yüksek iken; katı puanlayıcılar tarafından yapılan puanlamaların ortalaması gerçek puanlara kıyasla daha düşük olmaktadır (Wolfe, 2014).

2.1.3.2. Halo Etkisi

Halo etkisi ilk olarak 1907 yılında F.L. Wells tarafından açıklanmış ve 1920 yılında Thorndike tarafından isimlendirilmiştir (Karakuş, 2008). Thorndike (1920) halo etkisini "bireyi, genel olarak iyi ya da kötü şeklinde düşünme ve puanlamayı birey hakkındaki bu genel izlenime göre yapma eğilimi" şeklinde tanımlamıştır. Thorndike'ın halo etkisi kavramını alanyazına kazandırmasının ardından, halo etkisi farklı araştırmacılar tarafından değişik şekillerde tanımlanmıştır. Engelhard (2002) halo etkisini, puanlayıcının öğrencinin performansının kavramsal olarak farklı ve bağımsız yönlerini ayırt etmede başarısız olması ve bu özelliklere benzer puanlar vermesi olarak ifade etmiştir. Halo etkisine ilişkin benzer bir tanım Myford ve Wolfe (2004) tarafından yapılmıştır. Myford ve Wolfe'a (2004) göre, halo etkisi, puanlayıcının, öğrencinin performansını puanlarken, performansın kavramsal olarak farklı yönlerine aynı puanı verme eğiliminde olmasıdır. Bu iki tanımdan yola çıkan Eckes (2009), halo etkisini öğrencinin performansının kavramsal olarak farklı özelliklerinin puanlayıcı tarafından ayırt edilememesi ve bu farklı özelliklere oldukça benzer puanlar verilmesi şeklinde tanımlamıştır. Puanlama işlemine halo etkisinin karıştığına işaret eden farklı göstergeler bulunmaktadır. Ölçülen performansın farklı

boyutları arasında halo etkisi karışmadan elde edilmek istenen korelasyonlara kıyasla oldukça abartılı korelasyonlar, öğrencinin sınav puanlarındaki düşük varyans, yüksek puanlayıcı ve öğrenci etkileşimi, öğrencinin yeteneğini gösterebileceği bağımsız fırsatların sayısında azalma değerlendirme işlemine halo etkisinin karıştığına işaret eden göstergeler arasında yer almaktadır (Bechger, Maris ve Hsiao, 2010; Viswesvaran, Schmidt, ve Ones, 2005). Değerlendirme işlemine halo etkisinin karışması, öğrencinin puanları arasındaki korelasyonun artmasına neden olarak testin çok güvenilir olduğu izlenimini uyandırsa da; testin güvenirliğinin düşmesine sebep olan önemli bir faktördür (Bechger, Maris ve Hsiao, 2010). Tablo 2.2'de (Wolfe, 2014) iki boyutlu bir özellik için halo etkisine örnek teşkil eden puanlamalar sunulmuştur.

Tablo 2.2. İki boyutlu bir özellik için halo etkisine örnek teşkil eden puanlamalar

Öğrencilerin Sıra Numaraları	Gerçek Puan		Halo Etkisi Karışan Puanlamalar	
	Boyut 1	Boyut 2	Boyut 1	Boyut 2
1	0	2	0	1
2	1	1	1	0
3	1	0	2	1
4	2	1	1	2
5	2	2	2	2
6	2	3	2	3
7	2	4	3	3
8	3	2	2	4
9	3	4	3	3
10	4	3	4	3
İki Boyut Arasındaki Korelasyon Katsayıları	.51		.63	

Tablo 2.2'nin ilk sütununda, performansı ölçülen bireylerin sıra numaraları yer almaktadır. Sonraki sütunlarda ise sırasıyla, ölçülen özelliğin birinci ve ikinci boyutuna karşılık gelen gerçek puanlar ile halo etkisinin karıştığı puanlamalara yer verilmiştir. İki boyut arasındaki korelasyon gerçek puanlamalar için .51 iken; halo etkisinin karıştığı puanlamalar için .63'tür (Wolfe, 2014). Puanlamalara halo etkisi karışması durumunda, ölçülen özelliğin farklı boyutları arasındaki korelasyon katsayıları, boyutlar arasındaki gerçek korelasyon değerlerine kıyasla daha yüksek olmaktadır.

2.1.3.3. Merkeze Yönelme Etkisi

Merkeze yönelme etkisi, puanlama ölçeğinin orta kategorisinin fazla kullanımı olarak tanımlanmaktadır (Landy ve Farr, 1983; Myford ve Wolfe, 2003;

Wolfe ve McVay, 2010). Puanlama ölçeğinin orta kategorilerinin baskın olarak kullanılması, puanlayıcının uç değerleri kullanmaktan kaçındığını yansıtmaktadır. Bu durum değerlendirmede heterojenlik eksikliğini beraberinde getirmektedir. Buna bağlı olarak, aşırı tutarlı uyum istatistiklerinin elde edilmesi kaçınılmaz olmaktadır (Engelhard, 1994). Merkeze yönelme etkisi genellikle, çoklu bir derecelendirme ölçeğinin kullanılması ve puanlama işleminin bu konuda pek fazla tecrübesi olmayan deneyimsiz puanlayıcılar tarafından yapılması sonucunda ortaya çıkmaktadır (Baird, Hayes, Johnson, Johnson ve Lamprianou, 2013). Örneğin; ölçülen performansın puanlanması konusunda yeterli bilgisi olmayan bir puanlayıcı düşük ve yüksek performans düzeyindeki öğrencileri birbirinden ayırt edebilir. Ancak ortalama performans gösteren öğrencileri birbirinden ayırt etmede başarısız olabilir ve bu öğrenciler için ölçeğin orta kategorilerini daha fazla kullanabilir. Bunun sonucunda, performans puanlarına merkeze yönelme etkisi karışabilmektedir. Benzer şekilde, puanlama ölçeğinin kategorileri hakkında yeterli bilgisi olmayan bir puanlayıcı öğrencilerin performans düzeylerini ayırt etmede başarısız olabilir ve puanlama ölçeğinin orta noktasına başvurabilir. Puanlayıcının, puanlama ölçeğinin orta noktasına çok fazla başvurması ise merkeze yönelme etkisine yol açmaktır (Farrokhi, Esfandiari ve Vaez Dalili, 2011). Merkeze yönelme etkisi, puanlama ölçeğinde yer alan kategorilerin kullanım örüntüsü incelenerek belirlenebilir (Abu Kassim, 2007). Tablo 2.3'te (Wolfe, 2014), yedili derecelendirmeye sahip bir rubrik kullanılarak yapılan değerlendirme işleminde, merkeze yönelme etkisine örnek teşkil eden puanlamalar gösterilmiştir.

Tablo 2.3. Merkeze yönelme etkisine örnek teşkil eden puanlamalar

Gerçek Puanlar	0	1	1	2	2	2	3	3	3	3	3	4	4	4	5	5	6
Merkeze Yönelme Etkisi	2	2	1	2	2	3	3	3	3	3	3	3	4	4	4	4	3

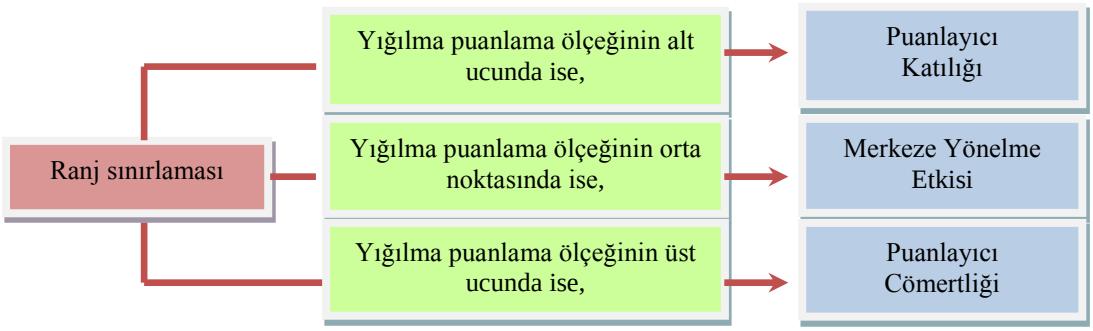
Tablo 2.3'ün ilk satırında gerçek puanlamalar yer alırken; ikinci satırında merkeze yönelme etkisinin karıştığı puanlamalar bulunmaktadır. Tablo 2.3'te de görüleceği üzere; merkeze yönelme etkisinin karıştığı puanlamalarda, puanlama ölçeğinin alt ve üst ucundan uzak durulup, daha çok orta kategorileri kullanılmıştır.

2.1.3.4. Ranj Sınırlaması

Ranj sınırlaması, merkeze yönelme etkisine benzemektedir (Engelhard, 1994; Engelhard ve Stone, 1998). Hem merkeze yönelme etkisi hem de ranj sınırlaması

yüksek derece uyumlu puanlar elde etme baskısından kaynaklanmaktadır. Ancak merkeze yönelme etkisinde, yığılma puanlama ölçeğinin orta noktasında olurken; ranj sınırlaması puanlama ölçeğinin herhangi bir noktasında olabilmektedir (Lottridge, Schulz ve Mitzel). Buna göre, ranj sınırlaması öğrencinin performansından bağımsız olarak, puanlayıcının puanlama ölçeğinin yalnızca belli kategorilerini kullanma eğiliminde olması şeklinde ifade edilmektedir (Moore, 2009). Örneğin, bazı puanlayıcılar puanlama ölçeğinin alt ucunu aşırı kullanırken; diğer puanlayıcılar ölçeğin üst ucunu daha yoğun bir biçimde kullanabilmektedir. Gerek merkeze yönelme etkisi gerekse de ranj sınırlaması öğrencilerin performanslarının doğru bir şekilde ayrılması konusunda başarısızlığa neden olduğundan, değerlendirmelerin geçerliğini tehdit eden unsurlardır (Saal, Downey ve Lahey, 1980).

Sıralanan dört puanlayıcı etkisinden halo etkisi, puanlamada birden fazla boyutun veya görevin olup olmadığı ile ilgili olup (Fisicaro ve Lance, 1990) analitik bir derecelendirme ölçeğinin kullanıldığı durumlarda görülmektedir (Robb, Singer ve LeMahieu, 2011). Diğer üç puanlayıcı etkisi ise tek bir boyutta ya da tek bir görev düzeyinde kalmaktadır. Halo etkisi bu yönüyle diğer üç puanlayıcı etkisinden farklılık göstermektedir (Fisicaro ve Lance, 1990). Puanlayıcı katılığı ve cömertliği, merkeze yönelme etkisi ve ranj sınırlaması arasındaki ilişki; Saal, Downey ve Lahey (1980) tarafından açıklanmıştır. Bu açıklamaya göre, ranj sınırlaması performansa verilen puanların puanlama ölçeğinin herhangi bir noktasında yığılması anlamına gelmektedir. Yığılmanın olduğu bu nokta puanlama ölçeğinin üst ucunda ise puanlayıcı cömertliği, alt ucunda ise puanlayıcı katılığı ve orta noktasında ise merkeze yönelme etkisi söz konusu olmaktadır. Dolayısıyla, puanlayıcı katılığı ve cömertliği ile merkeze yönelme etkisi ranj sınırlamasının özel durumları olarak ifade edilmektedir.



Şekil 2.3. Ranj sınırlaması ile merkeze yönelme etkisi, puanlayıcı katılığı ve puanlayıcı cömertliği arasındaki ilişkiler

2.1.3.5. Tutarsızlık (Randomness/Inconsistency)

Tutarsızlık; puanlayıcılardan herhangi birinin puanlama ölçeğini diğer puanlayıcılara göre daha farklı yorumlaması ve kullanmasıdır (Myford ve Wolfe, 2004). Puanlayıcı katılığı ve cömertliği, merkeze yönelme etkisi, ranj sınırlaması ve halo etkisi puanlamalardaki sistematik hataları gösteren tek biçimli (uniform) puanlayıcı etkileridir. Öte yandan; tutarsızlık, tek biçimli olmayan bir puanlayıcı etkisi olup puanlamalarda tesadüfi değişkenliklere yol açmaktadır (Ramineni, 2008).

2.1.3.6. Yanlılık/Değişen Puanlayıcı Katılığı ve Cömertliği (Bias/Differential Rater Severity and Leniency)

Yanlılık, değerlendirme işleminin bir yönü ile ilgili olarak, puanlayıcının beklenmedik bir biçimde cömert ya da katı puanlamalar yapmasıdır (Knock, Read ve Randow, 2007). Bir başka deyişle, değerlendirme işlemini yaparken puanlayıcının cinsiyet, yaş, kültürel faktörler gibi çeşitli özelliklerine bağlı olarak bazı bireylere diğerlerine göre daha yüksek ya da daha düşük puanlar verme eğiliminde olması yanlılık etkisi şeklinde tanımlanmaktadır (Kumar, 2005).

Puanlayıcılardan kaynaklanan bu etkiler ölçülen yapı ile ilgisiz varyansa neden olmaktadır (Bachman, 2004; Messick, 1995; Weir, 2005). Bu nedenle, kabul edilebilir bir güvenirlik düzeyine ulaşılması, söz konusu puanlayıcı etkilerinin minimum düzeye indirilmesi ile mümkün olabilir (Knoch, Read, von Randow, 2007; Zewotir, 2012). Puanlayıcı etkisini minimum düzeye indirebilmek için puanlama sürecinde birden fazla puanlayıcının kullanılması (Ebel, 1951), puanlayıcı eğitimlerine başvurulması (Woehr, 1994) ve puanlamaların rubriklere dayalı olarak yapılması (Wolf ve Stevens, 2007) gibi farklı öneriler getirilmiştir.

2.1.4. Performans Değerlendirmeye Karışan Puanlayıcı Etkilerini Azaltmak için Başvurulabilecek Uygulamalar

2.1.4.1. Puanlama Sürecinde Birden Fazla Puanlayıcının Kullanılması

Her bir performansın birden çok puanlayıcı tarafından puanlanması ve farklı puanlayıcılar tarafından verilen puanların ortalamasının alınması puanlayıcı etkisini minimum düzeyde tutmak için getirilen önerilerden biridir (Ebel, 1951). Ancak bu öneri, iki endişeyi beraberinde getirmektedir. Bunlardan ilki, her bir öğrencinin performansının birden fazla puanlayıcı tarafından puanlanmasının pratik bir

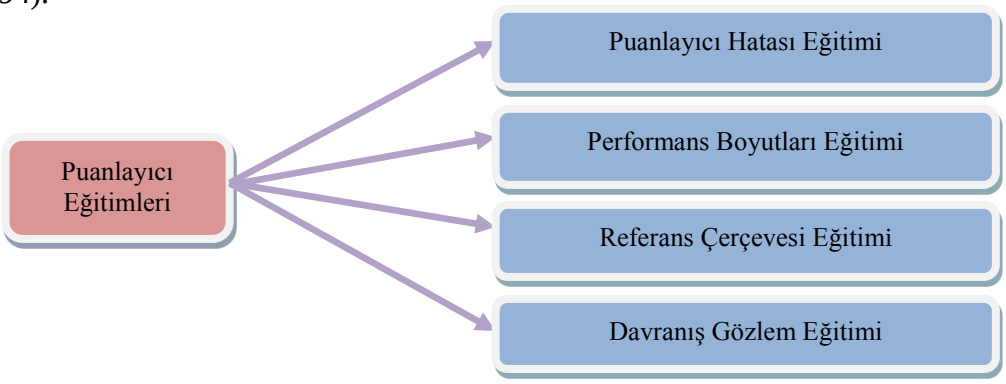
uygulama olmamasıdır. İkinci olarak, her bir öğrencinin yeteneğini güvenilir bir şekilde kestirmek için gerekli olan puanlayıcı sayısının ne olduğu sorusunu yanıtlamak mümkün değildir (Sudweeks, Reeveb ve Bradshawc, 2005).

2.1.4.2. Puanlayıcı Eğitimleri

Puanlayıcı etkisini minimum düzeyde tutmak için önerilen bir diğer yaklaşım puanlayıcı eğitimleridir (Lievens, 2001; Woehr, 1994). Puanlayıcı eğitimleri ile değerlendirmede kullanılacak puanlama ölçeklerinin puanlayıcılara tanıtılması (Fahim ve Bijani, 2011), değerlendirme işlemine yönelik örnek uygulamaların yaptırılması (Cellar, Curtis, Kohlepp, Poczapski ve Mohiuddin, 1989) ve böylelikle puanlayıcılar arasında ortak bir anlayış oluşturulması amaçlanmaktadır (Alderson, Clapham ve Wall, 1995). Buna göre, puanlayıcı eğitimlerinin ilk basamağını puanlayıcıların; test formatı, performans görevleri ve puanlama ölçütleri hakkında bilgilendirilmesi oluşturmaktadır. Bir başka deyişle, *i)* ölçülen yapı, *ii)* ölçülmesi amaçlanan performansın düzeyi/düzeyleri, *iii)* her bir performans düzeyindeki yapıyı temsil eden ölçütler, *iv)* puanlama ölçeğinin kategorileri ve *v)* katılımcıların yanıtladığı maddelerin ya da görevlerin zorluk seviyesi açısından puanlayıcılar arasında ortak bir anlayış oluşturulması puanlayıcı eğitimlerinin ilk adımıdır (Baird, Hayes, Johnson, Johnson, ve Lamprianou, 2013). Puanlayıcı eğitimlerinin ikinci adımını, puanlayıcılara birkaç performans üzerinde örnek uygulamalar yaptırılması oluşturmaktadır. Örnek uygulamalar ile puanlama ölçütleri açısından puanlayıcılar arasında ne ölçüde ortak bir anlayış olduğunun belirlenmesi hedeflenmektedir (Hamilton, Reddel ve Spratt, 2001). Örnek uygulamalar yapılırken, çeşitli yeterlilik düzeylerini temsil eden ve puanlama ölçeğinin uygulanması sırasında karşılaşılabilecek bazı tipik sorunları içeren performansların seçilmesine özen gösterilmelidir (Elder, Barkhuizen, Knoch ve Randow, 2007; Knock, Read ve Randow, 2007). Daha sonra, puanlayıcılar ile yaptıkları puanlamalar hakkında grup tartışmaları gerçekleştirilir. Grup tartışmaları puanlayıcıların hatalarını görebilmelerini ve bu hataların çözümüne yönelik neler yapabileceklerini fark etmelerini sağlar. Puanlayıcı eğitimlerinin son adımını ise; puanlayıcılara yaptıkları puanlamalar ile ilgili geri dönütlerin verilmesi oluşturmaktadır (Pulakos, 1991).

Puanlayıcı eğitimleri; puanlayıcı hatası eğitimleri (rater error training), performans boyutları eğitimi (performance dimension training), davranış gözlem eğitimi (behavioral observation training) ve referans çerçevesi eğitimi (frame of

referans training) olmak üzere dört başlıkta incelenmektedir (Woehr ve Huffuct, 1994).



Şekil 2.4. Puanlayıcı eğitimlerinin türleri

2.1.4.2.1. Puanlayıcı Hatası Eğitimleri

Puanlayıcı hatası eğitimlerinin amacı; puanlayıcı katılığı ve cömertliği, merkeze yönelme etkisi, halo etkisi, ranj sınırlaması, tutarsızlık ve yanlılık gibi puanlayıcı etkilerini minimum düzeye indirerek puanlamaların güvenirliğini arttırmaktır (Newman, Kinney ve Farr, 2005). Bu eğitim kapsamında, sıralanan puanlayıcı hataları ile ilgili olarak puanlayıcılara bilgi verilir. Daha sonra, bu puanlayıcı hatalarına yönelik örnekler sunulur (May, 2005; Woehr ve Huffuct, 1994). Söz gelimi, halo etkisinin minimum düzeyde tutulabilmesi için ilk olarak halo etkisinin tanımı verilip, hangi durumlarda halo etkisinin puanlamalara karışabileceği ifade edilir. Ardından, puanlama işlemine halo etkisinin karışmasını önleyecek uygulamalardan bahsedilir. Bir öğrenciyi değerlendirirken ölçülen performansın farklı boyutlarını aynı anda puanlamak yerine; ilk olarak tüm öğrencilerin performansının birinci boyutunun puanlanması, daha sonra tüm öğrencilerin performansının ikinci boyutunun puanlanması şeklinde bir yol izlenmesi, halo etkisini minimum düzeyde tutmak için başvurulacak uygulamalardan biri olabilir. Halo etkisine yönelik bu şekildeki bir eğitim, performansın farklı boyutlarının birbirinden ayırt edilememesi tehlikesini azaltabilir. Benzer şekilde, merkeze yönelme etkisinin puanlamalara karışmasını önlemek için önce bu hata puanlayıcılara tanıtılır. Daha sonra, puanlayıcılar, puanlama ölçeğinin yalnızca orta kategorilerini değil; diğer kategorilerini de kullanmaları konusunda uyarılır. Böylelikle, merkeze yönelme etkisinin minimum düzeyde tutulması amaçlanır. Özetle, puanlayıcı hatası eğitimleri ile değerlendiriciler, puanlamalar sırasında "ne yapmaları" ve "ne yapmamaları" gerektiği konusunda bilgilendirilir (Pulakos, 1991).

2.1.4.2.2. Performans Boyutları Eğitimi

Performansın farklı boyutlarının birbirinden ayırt edilebilmesi için puanlayıcılara performansın boyutları hakkında bilgi verilir (Newman, Kinney ve Farr, 2005). Bu sayede, puanlayıcıların performansın farklı boyutlarını birbirinden daha rahat bir biçimde ayırt etmesi sağlanabilir. Performans boyutları eğitimi özellikle, halo etkisini azaltmaya yönelik bir işlemdir. Ayrıca performans boyutları eğitimi ile değerlendirmede kullanılacak puanlama ölçeği puanlayıcılara tanıtılmaktadır (Cellar, Curtis, Kohlepp, Poczapski ve Mohiuddin, 1989).

2.1.4.2.3. Davranış Gözlem Eğitimi

Çoğu zaman puanlayıcılar performansa ilişkin gözlemlerini, değerlendirilecek performans dışındaki faktörlerin puanlamaya karışabildiği ortamlarda yapmak zorunda kalmaktadır (Noonan ve Sulsky, 2001). Dolayısıyla, puanlama sürecinin ilk basamağını oluşturan gözlem aşamasında, puanlayıcıların dikkatli gözlemler yapması ve gözlem sırasında oluşabilecek sistematik hatalardan kaçınması oldukça önemlidir (Roch ve O'Sullivan, 2003). Bunun için puanlayıcıların değerlendirdikleri bireyin performansını doğru bir şekilde gözleyebilecek yeterliliğe sahip olmaları gerekir (Noonan ve Sulsky, 2001). Davranış gözlem eğitimleri ile not alma, günlük tutma ve frekans hesaplama gibi (Newman, Kinney ve Farr, 2005; Woehr ve Huffuct, 1994) performans boyutlarıyla ilgili belirli davranışların gözlenmesini ve hatırlanmasını geliştirmeye yönelik becerilerin puanlayıcılara kazandırılması amaçlanır (Kozlowski, 2012; Labig ve Chye, 1996). Bu beceriler gözlem kayıtlarının başarılı bir biçimde tutulmasını sağlayarak değerlendirme işlemine karışabilecek puanlayıcı etkilerini azaltabilir (Antonioni ve Woehr, 2001).

2.1.4.2.4. Referans Çerçevesi Eğitimi

Performans standartları eğitimi olarak da adlandırılan bu eğitim ile performans tanımlanır ve performans boyutlarına ilişkin açıklamalar sunulur. Ardından puanlayıcılara başarılı, orta ve başarısız performans düzeylerini temsil eden performans örnekleri sunulur (Newman, Kinney ve Farr, 2005), puanlayıcılardan bu performans örneklerini puanlamaları istenir ve yapılan puanlamalar uzman puanlayıcılar tarafından yapılan puanlamalar ile karşılaştırılır (May, 2005; Sinclair, 2000). Son olarak da; puanlayıcılara puanlamaları ile ilgili ayrıntılı geri dönütler verilir (McIntyre, Smith ve Hassett, 1984). Sıralanan bu

işlemler sayesinde, puanlayıcılar arasında ortak bir referans çerçevesi oluşturulur (Bernardin ve Buckly, 1981). Böylelikle, bireylerin performanslarının farklı puanlayıcılar tarafından benzer bir biçimde değerlendirilmesi hedeflenir (McIntyre, Smith ve Hassett, 1984). Referans çerçevesi eğitiminde, değerlendirilecek performansın çok boyutlu olduğu esas alınmaktadır (Schmitt, 2012; Selden, Sherrier ve Wooters, 2012). Bu nedenle, referans çerçevesi eğitimi verilmeden önce puanlayıcının performans boyutları eğitimini almış olması gerekir (Sinclair, 2000). Ayrıca, referans çerçevesi eğitimi verilmeden önce, puanlayıcılar davranış gözlem eğitimi almış olmalıdır. Çünkü referans çerçevesi, performans değerlendirmenin karar alma sürecini oluşturmakta ve gözlenen davranışların sınıflandırılması ile değerlendirilmesine odaklanmaktadır. Bu aşamada doğru değerlendirmeler yapılabilmesi, performans değerlendirmenin ilk kısmını oluşturan davranışların gözlenmesi aşamasında dikkatli gözlemler yapılmış olmasına bağlıdır. Bu durum, referans çerçevesi eğitiminden önce davranış gözlem eğitiminin verilmesini gerekli kılmaktadır (Roch ve O'Sullivan, 2003).

Gröüldüğü üzere, alanyazında puanlayıcı eğitimlerine ilişkin farklı yaklaşımlar bulunmaktadır. Bu yaklaşımlardan hangisi kullanılırsa kullanılsın, puanlayıcı eğitimlerinden sonra, puanlayıcılar arasındaki farkın pratikte önemsiz olacak kadar küçük; puanlayıcılar arası güvenirliğin ise mümkün olduğu kadar yüksek olması arzu edilmektedir. Araştırmalar, puanlayıcı eğitimlerinin puanlayıcı tutarlılığını arttırırken; puanlamadaki aşırı farklılıkları ve bir performansa yönelik olarak puanlayıcılar tarafından ortaya konulan bireysel yanlılıkları azalttığını ortaya koymaktadır (Bernardin ve Walter, 1977; Fahim ve Bijani, 2011; McIntyre, 1993; Saito, 2008; Spool, 1978; Şahinkarakaş, 2010; Weigle, 1998; Wigglesworth, 1993; Woehr ve Huffuct, 1994). Ancak puanlamadaki farklılıkların ve bireysel yanlılıkların tamamen giderilmesi konusunda puanlayıcı eğitimleri başarılı olamamaktadır (Galati, 2007). Ayrıca araştırmalar, puanlayıcı eğitimlerinin etkisinin oldukça kısa süreli olduğunu ortaya koymuştur (Farrokhi, Esfandiari ve Vaez Dalili, 2011). Örneğin; Lumley ve McNamara (1995) tarafından yapılan araştırmada, puanlayıcı eğitiminin hemen ardından ve puanlayıcı eğitiminden bir ay sonra değerlendiricilerin puanlamaları arasında büyük farklılıkların olduğu belirlenmiştir. Benzer şekilde, Congdon ve McQueen (2000) tarafından okuma-yazma ve aritmetik testi ile ilgili olarak yapılan araştırmada, puanlayıcı eğitiminin üzerinden geçen zamana bağlı olarak puanlayıcıların cömertlik ve katılıklarında önemli değişiklikler olduğu

kaydedilmiştir. Bu durum puanlayıcı eğitiminin etkili olabilmesi için sürekli eğitime ihtiyaç duyulduğunu göstermektedir (Wang, 2010). Ancak bu şekildeki bir uygulama işlevsel görünmemektedir.

2.1.4.3. Rubrikler

Performansa dayalı değerlendirmelerde, puanlayıcı içi ve puanlayıcılar arası güvenirliği arttırmanın yollarından biri de değerlendirme sürecinde rubrik kullanılmasıdır (Dunbar, Brooks ve Miller, 2006). Rubrikler, 1990'lı yıllardan günümüze kadar farklı araştırmacılar tarafından değişik biçimlerde tanımlanmıştır. Aşağıdaki tabloda, rubriğe ilişkin farklı araştırmacılar tarafından yapılan bazı tanımlara yer verilmiştir.

Tablo 2.4. Farklı araştırmacılar tarafından rubriğe ilişkin yapılan tanımlar

Araştırmacı	Rubrik Tanımı
Brookhart (1999)	Bir performansa yönelik olarak öğrencilerin gösterdiği çaba ile ortaya koydukları ürünü değerlendirmede kullanılmak üzere geliştirilen açıklayıcı puanlama şemaları
Valenza (2000)	Bir performans açısından önemli olan davranış ve görevleri örneklendiren puanlama araçları
Venn (2000)	Bir performansın farklı düzeylerine ilişkin nitelikleri ve karakteristik özellikleri belirten puanlama ölçütleri
Mertler (2001)	Performans değerlendirmede kullanılan puanlama ölçekleri
Hafner ve Hafner (2003)	Okul öncesi eğitimden yükseköğretime kadar performansa dayalı değerlendirme çıktılarını ölçmek için kullanılan değerlendirme araçları
Jonsson ve Svingby (2007)	Otantik ya da karmaşık öğrenci görevlerinin puanlanmasında kullanılan puanlama araçları

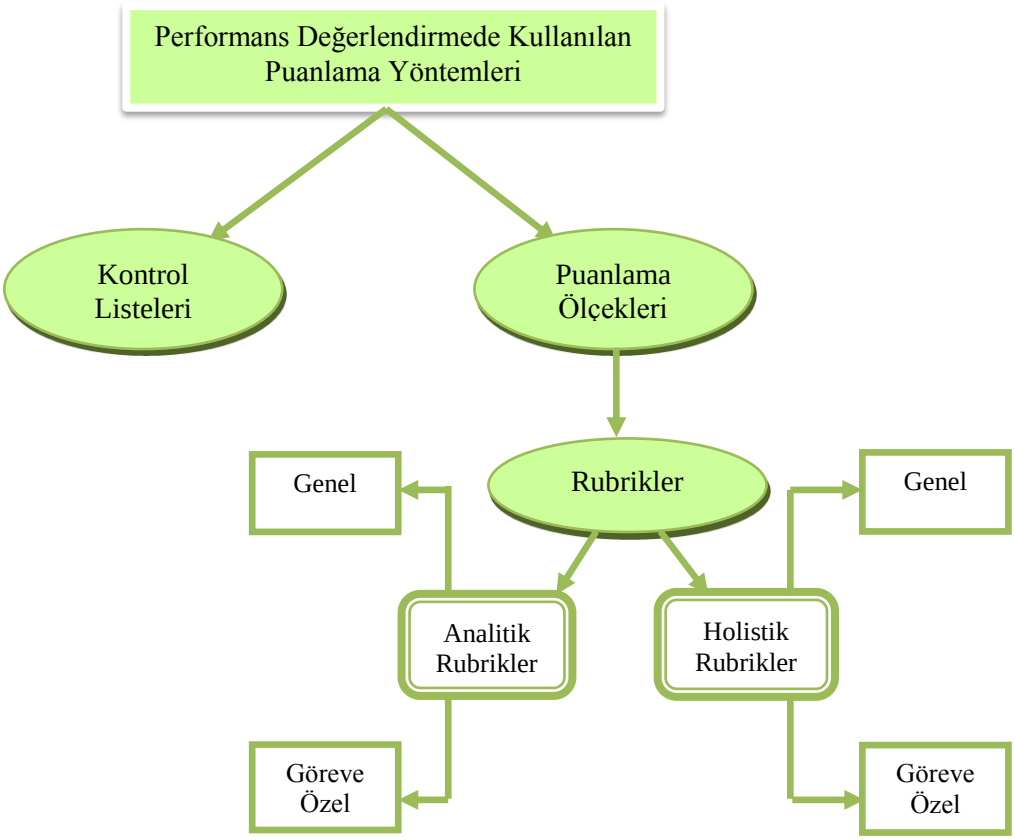
Tablo 2.4'te görüldüğü gibi, rubriklerin nasıl tanımlanması gerektiği konusunda araştırmacılar arasında tam bir uzlaşı bulunmamaktadır. Kan (2007), farklı araştırmacılar tarafından yapılan tanımlardan yola çıkarak rubriğe ilişkin genel bir çerçeve çizmeye çalışmıştır. Kan'a (2007) göre rubrik, performansın değişik düzeylerine ait karakteristik özellikler ile ölçütleri tanımlayan ve bu özellik ile ölçütler doğrultusunda performansa ilişkin yargıya varmada kullanılan bir puanlama rehberidir. Rubrikler, farklı araştırmacılar tarafından değişik biçimlerde tanımlansa da; değerlendirme sürecinde rubrik kullanımının pek çok yarar sağladığı konusunda araştırmacılar hem fikirdirler. Rubrikler; performansa ilişkin beklentileri (Howell, 2011; Roblyer ve Wiencke, 2010), performans değerlendirme sürecinde hangi davranışlara bakılacağını ve hangi özelliklerin önemli görüleceğini açıklığa

kavuşturarak (Hall ve Salmon, 2003; Airasian, 2005), öğretmen ve öğrencilerde ortak bir algı oluşturmaktadır (Arter ve McTighe, 2001; Busching, 1998; Perlman, 2003). Rubrikler bu yönüyle, öğrenme sürecinin gelişmesine katkıda bulunmaktadır (Andrade ve Du, 2005; Jonsson ve Svingby, 2007). Rubriklerde yer alan ölçütler, öğretmenin dersi planlamasını ve dersin kazanımlarına ulaşmaya katkıda bulunacak etkinlikler hazırlamasını kolaylaştırmaktadır. Diğer bir deyişle, rubrikte yer alan ölçütler öğrenme durumlarının düzenlenmesinde bir rehber olarak işlev görmektedir (Gunning, 2001; Halonen vd., 2003). Rubrikler ile kendi çalışmalarını değerlendirme fırsatı bulan öğrenciler (Oaklef, 2009), bu sayede güçlü ve zayıf noktalarının farkına varabilmektedir (Mertler, 2001). Rubrikler ayrıca, öğrencilerin güçlü ve zayıf olduğu noktalar hakkında öğretmen, öğrenci ve veliler arasında bir iletişim aracı olarak görev yapmaktadır (Hall ve Salmon, 2003). Rubriklerin en önemli avantajlarından biri de, öğrenci tarafından verilen yanıtların objektif bir biçimde değerlendirilemediği durumlarda, puanlamalar hakkında detaylı bilgiler ve örnek yanıtlar sunarak ölçme hatasını önemli ölçüde azaltması (Dunbar, Koretz ve Hoover, 1991; Penny, Johnson ve Gordon, 2000), öğrencilerin performansına ilişkin güvenilir (tutarlı) ve geçerli (doğru) değerlendirmeler yapma olanağı sunmasıdır (Pellegrino, Baxter ve Glaser, 1999). Değerlendirmenin amacına uygun olarak hazırlanan ve açık bir biçimde tanımlanan puanlama kategorilerinden oluşan bir rubrik, puanlama işleminin kim tarafından ve ne zaman yapıldığından bağımsız olarak gerçekleştirilmesini sağlar. Böylece, performans değerlendirmeye karışabilecek puanlayıcı etkilerinin minimum düzeyde tutulmasına yardımcı olur (Moskal ve Leydens, 2000; Purpura, 2004). Buna göre, rubriklerin puanlayıcı içi ve puanlayıcılar arası güvenirliği arttırdığı söylenebilir (Moskal ve Leydens, 2000).

2.1.4.3.1. Rubriklerin Uygun Bir Değerlendirme Tekniği Olduğu Durumlar

Rubrikler öğrencilerin çalışmalarını değerlendirmek amacıyla kullanılabilecek pek çok alternatif yöntemden biridir. Bazı durumlarda rubrikler yerine, kontrol listeleri de kullanılabilmektedir. Kontrol listeleri kullanılarak performansa ilişkin belirli ölçütlerin karşılanıp karşılanmadığı belirlenebilir. Ancak, kontrol listeleri ölçütlerin ne düzeyde karşılandığı konusunda bir bilgi vermemektedir. Bu nedenle, performans değerlendirmede kontrol listelerinin uygun bir tercih olduğu durumların sayısı oldukça sınırlıdır. Rubrikler ise, performansa ilişkin ölçütlerin ne düzeyde karşılandığının belirlenmesini sağlayan tanımlayıcı

ölçeklere dayalıdır (Moskal, 2000). Dolayısıyla kompozisyon, sözlü sunum, proje, portfolyo, yazılı ve sözlü iletişim becerileri (Reddy ve Andrade, 2010) gibi öğrencilerin gösterdikleri çabanın ve ortaya koydukları ürünün objektif olarak değerlendirilmediği durumlarda puanlama işleminin güvenirliğini artırmak için rubriklerden yararlanılmaktadır (Braun, 1988; Harik, 2008).



Şekil 2.5. Performans değerlendirmede kullanılan puanlama yöntemleri (Kutlu, Doğan ve Karakaya, 2010; Mertler, 2001)

2.1.4.3.2. Rubriği Oluşturan Öğeler

Rubrikler; değerlendirme ölçütleri, kalite tanımları ve puanlama stratejileri olmak üzere üç temel öğeden oluşmaktadır (Popham 1997). Değerlendirme ölçütleri öğrencinin çalışmasının kalitesini belirlemede puanlayıcı tarafından dikkate alınan faktörlerdir. Bu ölçütler, gösterge ya da kurallar listesi olarak da tanımlanmaktadır. Değerlendirme ölçütleri, performansa ilişkin önemli olduğu düşünülen işlemleri yansıtmakta (Parke, 2001) ve değerlendirilecek performansın boyutlarını tanımlamaktadır (Tiermer ve Simon, 2001). Değerlendirme ölçütleri, rubrik ile değerlendirilmek istenen yeteneğe bağlı olarak, rubrikten rubriğe farklılık göstermektedir. Örneğin; öğrencilerin kompozisyon yazma becerileri

değerlendirilirken, organizasyon, gramer, kelimelerin seçimi, destekleyici ayrıntılar gibi değerlendirme ölçütleri kullanılabilmektedir (Popham, 1997).

Rubriği oluşturan ikinci temel öğe, kalite tanımlarıdır (Reddy, 2010). Öğrencilerin başarılarını ya da yeterlilik düzeylerini belirleyebilmek için orta, iyi ya da mükemmel gibi başarı düzeyleri tanımlanır. Öğrencinin bu başarı düzeylerine ulaşmak için ne yapması gerektiği kalite tanımları ile açık bir biçimde ifade edilmektedir (Reddy ve Andrade, 2010). Örneğin, öğrencinin yazdığı kompozisyonun puanlanmasına yönelik bir değerlendirmede, gramer bir puanlama ölçütü olarak belirlenebilir. Bu durumda, "Öğrencinin en yüksek puanı alabilmesi için kompozisyonda hiçbir gramer hatası bulunmamalıdır" ifadesi bir kalite tanımı olabilir (Popham, 1997). Kalite tanımları, hem puanlama sırasında kullanılmak hem de öğrenciye geri dönütüler vermek üzere iyi ve kötü yanıtların birbirinden ayırt edilmesini sağlamaktadır (Reddy ve Andrade, 2010). Rubrikteki her bir düzey için ayrı bir kalite tanımı kullanılmalıdır. Söz gelimi; rubrikte değerlendirilecek performansa ilişkin dört ayrı düzey bulunuyorsa, bu düzeylerin her birine yönelik olarak ayrı bir kalite tanımının rubrikte yer alması gerekmektedir (Popham, 1997).

Rubriği oluşturan üçüncü temel öğe ise; puanlama stratejileridir. Puanlama stratejisi öğrencinin gösterdiği çaba ya da ortaya koyduğu ürün hakkında verilen kararların yorumlanması için bir ölçek kullanımını içermektedir (Reddy, 2010; Reddy ve Andrade, 2010). Kullanılacak bu ölçek, holistik (bütüncül) ya da analitik olabilmektedir. Holistik puanlama ölçeklerinde, ürün ya da süreç hakkındaki genel izlenime dayanarak tek bir puan verilmektedir. Performansı oluşturan bileşenler için ayrı ayrı puanlama yapılmamaktadır (Popham, 1997). Analitik bir puanlama ölçeğinin kullanılması durumunda ise, performansı ya da ürünü oluşturan her bir bileşen ayrı ayrı puanlanmaktadır (Reddy, 2010). Holistik ve analitik puanlama ölçekleri daha ayrıntılı olarak aşağıda açıklanmıştır.

2.1.4.3.3. Holistik ve Analitik Rubrikler

2.1.4.3.3.1. Holistik Rubrikler

Holistik rubriklerde, öğrencinin performansı ya da ortaya koyduğu ürün bir bütün olarak ele alınmakta (Nitko, 2004) ve performansın kalitesine ilişkin genel bir izlenim ile puanlama yapılmaktadır (Pomplun, Capps ve Sundbye, 1998). Nitko (2004), öğrencilerin farklı yanıtlar verme olasılığının yüksek olması ve tek bir doğru

cevabın bulunmaması halinde, holistik rubrik kullanılmasını önermektedir. Holistik rubrik kullanımının çeşitli avantajları ve dezavantajları vardır (Becker, 2010/2011). Holistik rubrikteki bütünsel yaklaşım göz önüne alındığında, bu rubrikler ürün ya da süreç hakkında geniş bir resim elde etme açısından yararlı görülmektedir (Reddy, 2010). Holistik rubriklerde performans bir defa incelenerek puanlama yapıldığından değerlendirme işlemi kısa sürede sonuçlanmaktadır (Mertler, 2001; Wiseman, 2012b). Holistik rubriklerin ekonomik ve pratik olması (Weigle, 2002) özellikle çok sayıda öğrencinin katıldığı yetenek ve yerleştirme testleri gibi geniş ölçekli uygulamalarda, sistem yöneticileri tarafından sıklıkla tercih edilmesine neden olmaktadır (Becker, 2010/2011; Fraser vd., 2005; Jonsson ve Svingby, 2007). Diğer taraftan, holistik rubrikler öğrenciler hakkında yeterince ayrıntılı bilgi elde etmeye imkân tanımamaktadır (Bargainnier, 2003). Holistik rubrikler, öğrencilerin güçlü ve zayıf olduğu noktalara ilişkin bilgi vermemekte (Arter ve Mctighe, 2001), ne kadar ve nasıl bir ek öğretim hizmetine ihtiyaç duyulduğu hakkında yeterince açıklayıcı olamamakta (Cohen, 1994) ve bir sonraki öğretim hedefinin belirlenmesine yardımcı olabilecek herhangi bir bilgi sunmamaktadır (Nelson ve Van Meter, 2007). Bu nedenle, holistik rubriklerin daha çok düzey belirlemeye yönelik değerlendirmelerde kullanılması önerilmektedir (Mertler, 2001) .

2.1.4.3.3.2. Analitik Rubrikler

Analitik rubrikler, çeşitli boyutlardan oluşan performansların puanlanmasında kullanılmaktadır (Weigle, 2002). Analitik rubriklerde, performansı oluşturan bileşenlerin her biri ayrı ayı puanlanır (Klein vd., 1998). Daha sonra, her bir bileşen için verilen puanlar toplanarak performansa ilişkin genel bir puan elde edilmektedir (Petkov ve Petkova, 2006). Nitko (2004), açık bir biçimde tanımlanmış bir yanıtın gerekli olduğu durumlarda analitik rubriklerin kullanılmasını önermektedir. Diğer bir ifadeyle; bir ya da iki kabul edilebilir yanıtın olduğu ve yaratıcılığın, öğrencinin performansının önemli bir özelliği olmadığı durumlarda analitik rubrikler tercih edilmelidir (Mertler, 2001). Analitik rubrikler, tek bir bütünsel puanın verildiği holistik rubriklere kıyasla öğrencinin performansı hakkında daha detaylı bilgi vermektedir. Ayrıca, performansın farklı boyutlarına verilen puanlara dayalı olarak öğrencilerin ölçülen özelik açısından güçlü ve zayıf yönlerini ortaya koymaya imkân tanımaktadır (Brown ve Hudson, 2002; Jonsson ve Svingby, 2007; Wiseman, 2012b). Öğrencinin, performansı hakkında ayrıntılı geri bildirimler almasını sağlayan

(Bacha, 2001; Carr, 2000; Nitko, 2004) bu özelliğine bağlı olarak, tanıma ve biçimlendirmeye yönelik değerlendirmelerde analitik rubrikler sıklıkla tercih edilmektedir (Becker, 2010/2011; Mertler, 2001; Reddy, 2010). Yine aynı özellik analitik rubrikleri sınıf ortamındaki değerlendirmeler için uygun bir seçenek haline getirmektedir (Jonsson ve Svingby, 2007). Analitik rubriklerin holistik rubriklere göre en önemli avantajlarından biri, analitik rubriklerin kullanıldığı değerlendirmelerde hem puanlayıcı içi hem de puanlayıcılar arası güvenirliğin daha yüksek olmasıdır (Knoch, 2009). Analitik rubriklerin sıralanan avantajlarının yanı sıra bir takım dezavantajları da bulunmaktadır. En önemli dezavantajlarından biri, performansın farklı boyutlarının birbirinden başarılı bir biçimde ayırt edilememesidir. Bu durum "halo etkisi" olarak tanımlanmaktadır (Myford ve Wolfe, 2004). Halo etkisini minimum düzeyde tutabilmek için performansın farklı boyutlarının arka arkaya puanlanması yerine; her bir boyutun ayrı bir zamanda incelenmesi önerilmektedir (Mertler, 2001). Örneğin, üç boyutlu bir yapıya sahip bir performans değerlendirilirken, ilk olarak tüm öğrencilerin performansın birinci boyutundan, daha sonra tüm öğrencilerin performansın ikinci boyutundan ve son olarak tüm öğrencilerin performansın üçüncü boyutundan aldıkları puanlar incelenebilir. Bu şekildeki bir puanlama sayesinde, bir boyuttan alınan puanların diğer boyutlardan alınan puanları etkilemesinin önüne geçilebilir. Analitik rubriklerin bir diğer dezavantajı, hem geliştirilmesinin hem de uygulanmasının zaman alıcı ve maliyetli olmasıdır (Barkaoui, 2008; Weigle, 2002; Wiseman, 2012b).

Goulden (1992,1994), holistik ve analitik rubrikler arasındaki farka değişik bir bakış açısı getirmiştir. Goulden'e göre, holistik ve analitik rubrikler, değerlendirilen performansın geneli ve parçaları arasındaki ilişkiler hakkında farklı varsayımlara sahiptir (Akt: Barkaoui, 2008). Analitik rubriklerde, performansın farklı bileşenlerine ait puanların toplanmasıyla performansın genelinin değerlendirilmiş olacağı varsayılmaktadır. Tersine holistik rubriklerde, performansın bir bütün olduğu ve bütünün kendini oluşturan bileşenlerin toplamına değil; ancak bu bileşenler ile bileşenler arasındaki ilişkilerin toplamına eşit olabileceği kabul edilmektedir (Goulden, 1992). Dolayısıyla, holistik rubriklerde performansın bütününün, performansı oluşturan boyutların farklı ağırlıkları ve bu boyutlar arasındaki sinerjinin etkisiyle oluştuğu varsayılmaktadır. Bu varsayımın bir sonucu olarak holistik rubriklerde, performansın farklı boyutlarının ayrı ayrı puanlanması yerine; bir bütün olarak değerlendirilmesi esas alınmaktadır (Goulden, 1994).

2.1.4.3.4. Genel ve Göreve Özel Rubrikler

Puanlama rubrikleri belirli bir göreve özel olarak tasarlanabilir ya da benzer performansların değerlendirilmesinde ortak bir rubrik kullanılabilir (Moskal, 2000; Moskal ve Leydens, 2000). Örneğin, tüm öğrencilerin sözel sunum becerilerini değerlendirmek için bazen tek bir rubrik yeterli olabilir. Ancak, öğrenciler tarafından yapılan sunumların her biri faklı bir tarihsel olaya dayanıyorsa ve değerlendirmenin amacı, öğrencilerin tarih bilgisini değerlendirmek ise genel bir rubrik öğrenciler tarafından yapılan farklı sunumları değerlendirmede yetersiz kalabilir. Tarihsel olaylar, hem bu olayları etkileyen faktörler hem de olayların sonuçları yönüyle farklılık göstermektedir. Öğrencilerin bu olaylarla ilgili kavramsal ve olgusal bilgisini değerlendirmek için her bir sunuma yönelik olarak ayrı bir rubriğin geliştirilmesi gerekli olabilir. Çünkü her bir sunumda farklı bir tarihsel olay anlatılmaktadır. Tek bir konuya yönelik olarak hazırlanan bu şekildeki rubrikler göreve özel rubrik olarak adlandırılmaktadır (Moskal, 2000). Buna göre, benzer performansları değerlendirmek için tek bir rubrik kullanılabiliyorsa bu tür rubrikler *genel rubrik* olarak ifade edilmektedir. Diğer taraftan, her bir performans için ayrı bir rubrik geliştiriliyor ve bir performansın değerlendirilmesinde kullanılan bir rubrik bir başka performansın değerlendirilmesinde kullanılamıyorsa bu tür rubrikler *göreve özel rubrikler* olarak adlandırılmaktadır (Arter ve McTighe, 2001). Değerlendirilecek performans sayısının çok olduğu durumlarda genel rubriklerin kullanılabileceği ifade edilmektedir (Riddle ve Smith, 2008). Bununla birlikte, Nitko (2004) genel rubrikler yerine göreve özel rubriklerin kullanılmasını daha doğru bir tercih olarak nitelendirmektedir. Göreve özel rubrikler puanlama işleminin daha güvenilir olmasını sağlamaktadır. Ayrıca birçok durumda ölçülmek istenen özellik genel bir rubrik ile puanlanabilecek kadar basit bir yapıda değildir (Kutlu, Doğan ve Karakaya, 2010). Rubrikler genel ya da göreve özel hazırlanabileceği gibi, hem genel hem de göreve özel bileşenleri içerek şekilde de tasarlanabilir. Örneğin, öğrenciler tarafından yapılan sunumlarda, öğrencinin sözel sunum becerisi ve sunduğu tarihsel olayla ilgili bilgisi değerlendirilmek istenebilir. Bu durumda, öğrencilerin sözel sunum becerilerini değerlendirmek için genel bir rubrik ve sundukları tarihsel olaylarla ilgili bilgilerini değerlendirmek için ise göreve özel bir rubrik tasarlanabilir (Moskal, 2000).

2.1.4.3.5. Rubrik Geliştirme Basamakları

Mertler (2001), farklı araştırmacılar (Airasian, 2005; Montgomery, 2001; Nitko, 2004; Tombari & Borich, 1999) tarafından önerilen basamaklardan yola çıkarak rubrik geliştirme sürecini aşağıdaki şekilde açıklamıştır.

Adım 1: *Performans görevinin hitap ettiği öğrenme amaçlarının yeniden incelenmesi:* Bu işlem, eğitimin amaçları ve gerçek öğretim durumları ile puanlama rehberinin eşleştirilmesini sağlar.

Adım 2: *Öğrencilerin ürünlerinde, performanslarında ve öğrenme sürecindeki davranışlarında gözlenmek istenen ve istenmeyen özelliklerin tanımlanması:* Öğrencide aranan davranış, beceri ve özelliklerin yanı sıra görülmesi istenmeyen yaygın hataların belirtilmesi.

Adım 3: *Her bir gözlenen davranışı anlatan özelliklerin belirlenmesi*: İkinci adımda tanımlanan her bir davranış için ortalama, ortalamanın altı ve ortalamanın üstündeki performansların açıklanması.

Adım 4a: *Holistik rubrikler için, iyi ve kötü performansa ilişkin her bir davranışı tanıma dâhil eden ayrıntılı açıklamalar yazılması:* Gözlenen tüm davranışlar için yüksek ve düşük performans düzeylerinin tanımlanması.

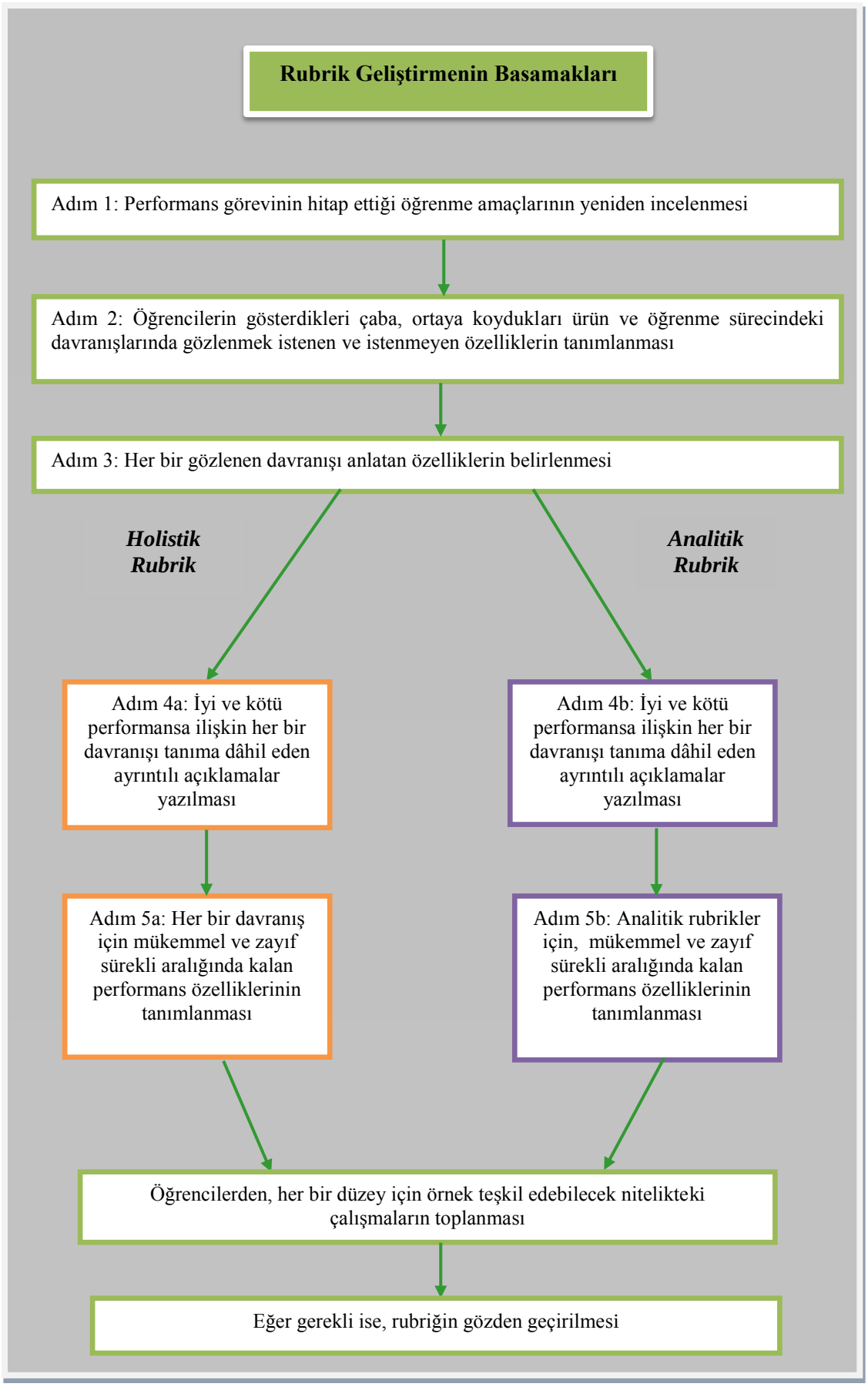
Adım 4b: *Analitik rubrikler için, gözlenen her bir davranışa yönelik olarak iyi ve kötü çalışmalara ait tanımlayıcı açıklamaların özet bir biçimde yazılması*: Her bir gözlenen davranış için yüksek ve düşük performans düzeylerinin ayrı ayrı tanımlanması.

Adım 5a: *Holistik rubrikler için mükemmel ve zayıf sürekli aralığında kalan performans özelliklerinin tanımlanması:* Performansa ilişkin mükemmel ve zayıf aralığındaki tüm düzeylerin tanımlanması.

Adım 5b: *Analitik rubrikler için mükemmel ve zayıf sürekli aralığında kalan performans özelliklerinin tanımlanması:* Performansa ilişkin mükemmel ve zayıf aralığındaki tüm düzeylerin her bir gözlenen davranış için ayrı ayrı tanımlanması.

Adım 6: *Öğrencilerden, her bir düzey için örnek teşkil edebilecek nitelikteki çalışmaların toplanması:* Bu çalışma örnekleri birer puanlama ölçütü olarak işlev görecektir.

Adım 7: *Eğer gerekli ise rubriğin gözden geçirilmesi:* Rubriğin etkililiğinin belirlenmesi ve sonraki uygulamalar için gözden geçirilmesi.



Şekil 2.6. Analitik ve holistik rubrik geliştirme basamakları

Hangi amaçla kullanılacağı, analitik ve holistik puanlama yöntemlerinden hangisinin tercih edileceği ve geliştirilecek rubriğin genel mi yoksa göreve özel mi olacağı rubrik geliştirme sürecinde izlenecek adımları etkileyen başlıca faktörlerdir (Kutlu, Doğan ve Karakaya, 2010). Bununla birlikte, rubrik geliştirme sürecini etkileyen başka faktörler de bulunmaktadır. Bu faktörlerden biri, rubriğin herhangi bir taksonomi temele alınmadan mı; yoksa bir taksonomiye dayalı olarak mı geliştirileceğidir. Rubrikler bazen herhangi bir taksonomi temele alınmadan geliştirilmektedir. Herhangi bir taksonomi temele alınmadan geliştirilen rubrikler için bu çalışmada *standart rubrik* ifadesi kullanılmaktadır. Standart rubriklerde, öğrenme çıktılarının değerlendirilmesinde kullanılacak ölçütler herhangi bir taksonomi göz önünde bulundurulmadan belirlenmekte ve *yetersiz*, *geliştirilmesi gerek*, *kabul edilebilir*, *iyi* ve *çok iyi* gibi derecelendirmeler esas alınmaktadır. Sorunun çözümünde takip edilen işlem basamakları, cevabın doğruluğu, çözüme yönelik olarak yapılan açıklamaların yeterliliği ve anlaşılırlığı dikkate alınarak öğrencinin verdiği cevabın rubriğin hangi düzeyine karşılık geldiği belirlenmektedir. Örneğin; açık uçlu bir matematik problemi için öğrencinin hem cevabının hem de çözüm yolunun hatalı olduğu yanıtlar *yetersiz* düzeyine; çözüm yolunun doğru ancak yapılan işlemlerin ve verilen cevabın hatalı olduğu yanıtlar *geliştirilmesi gerek* düzeyine; cevabın doğru, çözümde takip edilen basamakların açık ve anlaşılır olduğu yanıtlar ise *çok iyi* düzeyine karşılık gelebilir. Standart rubriklere örnek teşkil etmesi bakımından, MEB (2007) tarafından yayınlanan matematik öğretmen kılavuz kitabında yer alan ve öğrencilerin matematik problemlerini çözme becerilerini ölçmeye yönelik holistik bir rubrik Tablo 2.5'te sunulmuştur.

Tablo 2.5. Matematik problem çözme becerisi için standart rubrik örneği

	ÖLÇÜTLER	PUAN
1 Puan	Çalışma aşağıdaki özellikleri taşıyorsa bu puan verilecektir.	
	-Hiçbir çalışma yapılmamışsa -Sadece yanlış sonuç yazılmışsa -Problemdeki veriler sadece kopyalanmışsa veya problemi anlama izleri yoksa	
2 Puan	Çalışma aşağıdaki özellikleri taşıyorsa bu puan verilecektir.	
	-Problemin alt amaçlarından sadece biri üzerinde çalışılmış ve sonuçlandırılamamışsa -Çözümü bulmaya yönelik başlangıç yapılmış ancak bu başlangıç doğru cevabı bulmaya yeterli olmamışsa -Uygun olmayan strateji ile başlangıç yapılmışsa veya problem bu strateji ile çözülmeye çalışılmış fakat sonuçlandırılamamışsa	
3 Puan	Çalışma aşağıdaki özellikleri taşıyorsa bu puan verilecektir.	
	-Problem anlaşılmışsa ve uygun olmayan strateji ile başlangıç yapıldığı için yanlış sonuca ulaşılmışsa -Doğru sonuç bulunmuş ancak sonuç anlaşılmıyorsa -Sadece doğru sonuç varsa -Sadece problemin alt amaçlarından birinin çözümü doğru ise -Uygun strateji ile sadece başlangıç yapılmışsa -Uygun strateji seçilmiş ancak yanlış uygulanmışsa	
4 Puan	Çalışma aşağıdaki özellikleri taşıyorsa bu puan verilecektir.	
	-Problemi yanlış veya kısmen anlaşıldığı için uygun strateji kullanılmasına karşın yanlış sonuca ulaşıldıysa -Uygun strateji kullanılırken anlaşılmayan nedenlerden dolayı yanlış sonuca ulaşılmışsa -Uygun stratejinin kullanıldığı anlaşılmamasına karşın doğru cevap verilmişse -Uygun strateji uygulanmış fakat sonuç yazılmamışsa	
5 Puan	Çalışma aşağıdaki özellikleri taşıyorsa bu puan verilecektir.	
	-Uygun stratejiyi kullanılırken hata yapılmışsa ve bu hata problem anlaşılmadığı veya kavram yanılgısı olduğu için değilse -Uygun strateji kullanılmış ve doğru sonuca ulaşılmışsa	

Kimi zaman ise, rubrikte kullanılan değerlendirme ölçütlerinin belirlenmesinde, Bloom taksonomisi (Richardson, Ertmer ve Lehman, 2007; Chan, Tsui, Mandy ve Hong, 2002), yansıtıcı düşünme modeli (Chan vd., 2002) ve SOLO taksonomisi (Killen ve Hatting, 2004) gibi farklı taksonomilerden yararlanılmaktadır (Chan vd., 2002). Özellikle SOLO taksonomisine dayalı rubrikler açık uçlu soruları puanlamak için birçok farklı eğitim kademesinde ve birçok farklı derste sıklıkla kullanılmaktadır (Slack, Beer, Armitt ve Green, 2003).

2.1.5. SOLO Taksonomisi

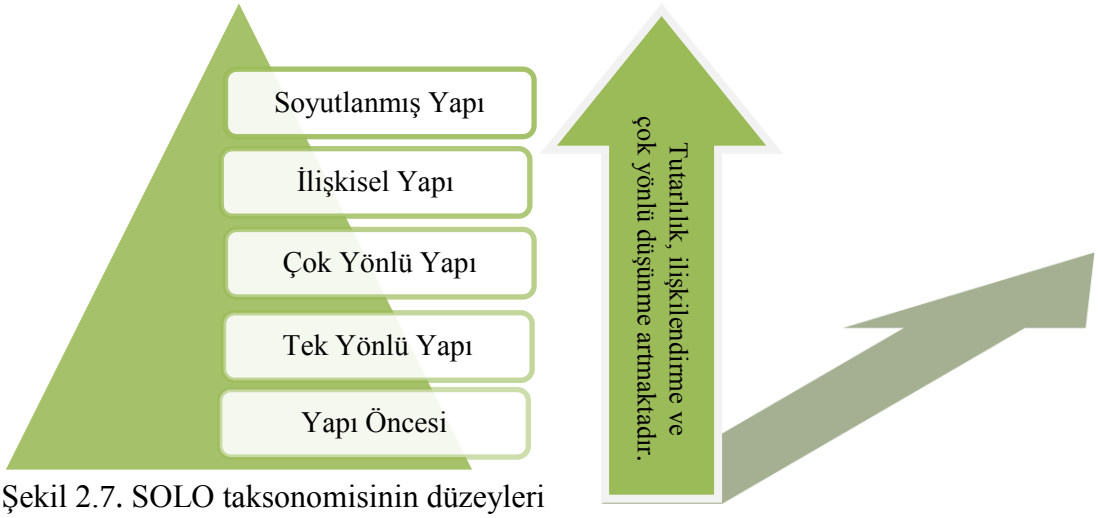
SOLO (Structure of the Observed Learning Outcome) taksonomisi, gözlenebilir öğrenme çıktılarının yapısını açıklamak üzere Biggs ve Collis (1982) tarafından geliştirilmiştir (Mohd Nor ve Idris, 2010). Taksonominin geliştirilmesinde Piaget'in bilişsel gelişim teorisi temele alınmıştır (Minogue ve Jones). Dolayısıyla

SOLO taksonomisi ile Piaget'in bilişsel gelişim teorisi arasında bir takım benzerlikler bulunmaktadır.

SOLO taksonomisinin; duyusal motor (sensori motor) düzeyi, imgesel (ikonic) düzey, somut sembolik (concrete sembolic) düzey, soyut (formal) düzey ve soyut sonrası (post formal) düzey olmak üzere Piaget'in bilişsel gelişim basamaklarına karşılık gelen beş düşünce evresinden oluşması iki model arasındaki benzerliklerden biridir (Pegg ve Tall, 2005). Hem SOLO taksonomisinin hem de Piaget'in bilişsel gelişim teorisinin hiyerarşik bir yapıya sahip düzeylerden oluşması iki model arasındaki bir diğer benzerliktir (Hattie ve Purdie, 1998). Öğrencinin herhangi bir kavram ile ilgili olarak, SOLO taksonomisinin belirli bir düzeyine ulaşması, bulunduğu seviyenin gerektirdiği işlemleri yapabilecek yeterlilikte olduğu anlamına gelmektedir (Slack vd., 2003). SOLO taksonomisi, bu yönüyle de Piaget'in bilişsel gelişim teorisi ile benzerlik göstermektedir. SOLO taksonomisi, bazı özellikleri açısından ise Piaget'in bilişsel gelişim teorisinden ayrılmaktadır. Piaget'in bilişsel gelişim teorisine göre, herhangi bir konunun bireye öğretilebilmesi için bireyin bilişsel açıdan söz konusu konuyu öğrenebilecek hazır bulunuşluk düzeyine ulaşmış olması gerekir. Öte yandan, SOLO taksonomisinde, hazır bulunuşluk kavramı daha farklı bir biçimde ele alınmaktadır. SOLO taksonomisinde ilgi odağı öğrencinin bulunduğu zihinsel gelişim döneminden, belirli bir görev ile ilgili olarak ortaya koyduğu performansın kalitesine ya da belirli bir soruya verdiği yanıtın niteliğine kaymaktadır (Hattie ve Purdie, 1998). Buna göre, SOLO taksonomisinin odak noktası öğrencinin yaşına karşılık gelen bilişsel gelişim dönemi değil; öğrenme sürecinde ortaya koyduğu performanstır. Bu durum, SOLO taksonomisi ile Piaget'in bilişsel gelişim teorisi arasındaki en önemli farklılıklardan biridir (Gilbert, 2004). SOLO taksonomisinin Piaget'in bilişsel gelişim teorisinden bir diğer farkı da, SOLO taksonomisinin belirli bir zamanda ve belirli bir performansa yönelik olarak tanımlanmasıdır (Hawkins ve Hedberg, 1986; Pegg ve Tall, 2005). SOLO taksonomisinde, öğrencinin bulunduğu düzeyin, konuya ve içeriğe göre farklılık gösterebileceği esas alınmaktadır (Ramsden, 2002). Piaget'in bilişsel gelişim teorisine göre ise, öğrencinin bulunduğu bilişsel gelişim basamağı konu alanından ve öğrenme içeriğinden bağımsızdır.

2.1.5.1. SOLO Taksonomisinin Düzeyleri

SOLO taksonomisi; yapı öncesi, tek yönlü yapı, çok yönlü yapı, ilişkisel yapı ve soyutlanmış yapı olmak üzere beş düzeyli bir yapıya sahiptir (Wong, 2007). Bu beş düzey, öğrenci herhangi bir soruyu yanıtlarken cevabını beş farklı şekilde yapılandırabileceğini göstermektedir (Lucas ve Mladenovic, 2008). Her bir düzeyin özellikleri aşağıda açıklanmıştır.



Şekil 2.7. SOLO taksonomisinin düzeyleri

2.1.5.1.1. Yapı Öncesi

Bu seviye, SOLO taksonomisinin en alt düzeyidir. Bu düzeyde, öğrenciler üzerinde çalıştıkları konuyu hiç anlamamakta ya da çok az anlamaktadır (Biggs, 1995). Üzerinde çalışılan durumun, cevapla ilişkisi olmayan yönleri öğrencinin dikkatini dağıtıp onu yanlış yönlendirmektedir (Burnett, 1999). Öğrencinin verdiği cevabın çözmeye çalıştığı problemle pek bir ilgisi yoktur (Brabrand ve Dahl, 2009). Öğrencinin ileri sürdüğü fikirler problemin çözümü için herhangi bir yarar sağlamamaktadır (Leung, 2000). Öğrenci problemi anlamadan daha önceki tecrübelerine dayanarak hızlı bir biçimde cevaplamaya çalışır (Jaskari, 2013). Bu durum öğrencinin verdiği yanıtın kişisel ya da öznel olmasına neden olmaktadır. Dolayısıyla, öğrenci gerçekleştirmesi beklenen görevi uygun bir biçimde yerine getirememektedir (Lian ve Yew, 2012).

2.1.5.1.2. Tek Yönlü Yapı

Bu düzeyde öğrenci sunulan konuyu ya da problemi dar ve yüzeysel bir bakış açısı ile ele almaktadır (Leung, 2000). Öğrenciler bu düzeyde; çalışılan konu ile ilgili terminolojiyi kullanabilir, konu ile ilgili açıklamalar yapabilir, basit talimatları ve

algoritmaları gerçekleştirebilirler (Brabrand ve Dahl, 2009). Bu düzeyde, öğrencilerin problemin tek bir yönüne odaklanması söz konusudur (Groth ve Bergner, 2006). Öğrenciler, odaklandıkları parçanın bütün içindeki yerini kavrayamamakta ve bu parçanın diğer parçalar ile ilişkisini kuramamaktadır. Bu nedenle, öğrencinin probleme verdiği cevapta birtakım tutarsızlıklar olabilmektedir. Öğrenci problemin çözümüne yönelik olarak oldukça basit bir çözüm ileri sürmektedir. Öğrencinin soruya verdiği cevaplar tam olmaktan uzaktır (Leung, 2000). Açıklamak, tanımlamak, ezberlemek, basit bir işlemi uygulamak, adlandırmak, sıralamak ve saymak tek yönlü yapı düzeyinin gösterge fiilleri arasında yer almaktadır (Brabrand, 2008; Brabrand ve Dahl, 2009).

2.1.5.1.3. Çok Yönlü Yapı

Bu seviyede, konu ya da probleme ilişkin temel noktalar göz önünde bulundurulur. Öğrenci üzerinde çalışılan konunun çeşitli yönlerini görebilir. Ancak bu yönler birbirlerinden bağımsız olarak kabul edilmektedir. Konunun farklı yönleri arasında bağlantılar kurulması söz konusu değildir (Padiotis ve Mikropoulos, 2010). Bir başka deyişle, öğrenci konu hakkında bir şeyler bilmekte ancak bildiklerini basit bir biçimde listelemekten öteye geçememektedir (Jimoyiannis, 2011). Öğrencinin problemin çözümüne yönelik olarak yaptığı açıklamalar ve ileri sürdüğü fikirler birçok bileşen içermektedir. Bununla birlikte, ileri sürülen fikirlerin organizasyonu zayıftır. Üretilen fikirler tutarlı bir biçimde bir araya getirilememektedir. Dolayısıyla, cevaplama sırasında kullanılan bazı ifadeler bir başkası ile tutarsız olabilmektedir (Leung, 2000). Bu düzeyde açıklayıcı bir yaklaşım mevcuttur. Ancak, neden-sonuç ilişkileri kurulamamaktadır (Kanuka, 2011). Örneğin, bir öğrenci iki yapı arasındaki benzerlikleri ve farklılıkları sıralayabilmesine rağmen, bu yapılar arasında nasıl bir ilişki bulunduğu hakkında bir fikir sahibi olmayabilir (Tan, Tan ve Chua, 2008). Lister, Simon, Thompson, Whalley ve Prasad (2006) bu düzeyi, öğrencilerin ağaçları görebildiği, ancak ormanı göremediği bir düzey olarak ifade etmektedir. Birleştirmek, sınıflandırmak, numaralandırmak, listelemek, tanımlamak, metaforik konuşmak, planlamak, algoritmaları ve yöntemleri uygulamak çok yönlü yapı düzeyinin gösterge fiilleri arasında yer almaktadır (Brabrand, 2008; Brabrand ve Dahl, 2009).

2.5.5.1.4. İlişkisel Yapı

Bu düzeyde öğrenci olayın çeşitli yönlerini görmekle kalmaz. Aynı zamanda, üzerinde çalışılan konu ya da probleme ilişkin bileşenleri tutarlı bir bütün oluşturacak şekilde birbirleriyle ilişkilendirebilir (Weyers, 2006). Bu ilişkilendirme sürecinde, her bir parça genel anlama katkıda bulunmaktadır (Kanuka, 2011). Ulaşılan sonuç, benzer bir duruma ya da probleme uygulanır (Lucas ve Mladenovic, 2008). Bu düzeyde genellemeler yapabilmektedir. Ancak yapılan genellemeler, mevcut bilgiler ile sınırlıdır. Öğrencinin, mevcut bilgilerin ötesinde bir sonuca ulaşması söz konusu değildir (Lake, 1999). Lister vd. (2006) bu düzeyi öğrencilerin ağaçları görmekle kalmayıp ormanı da görebildiği bir düzey olarak nitelendirmektedir. Analiz etmek, karşılaştırmak, birleştirmek, ilişkilendirmek, sebep ve sonuçları açıklamak, verilen bir teoriyi ilgili alana uygulamak (Brabrand, 2008; Brabrand ve Dahl, 2009), bir işlemi tersine çevirmek ilişkisel yapı düzeyinin gösterge fiilleri arasında bulunmaktadır (Joy ve Wilmot, 2008).

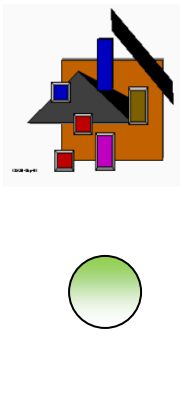
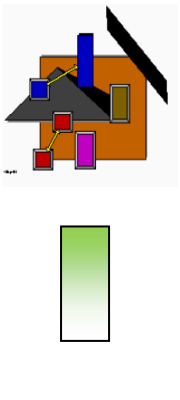
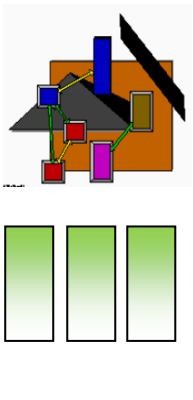
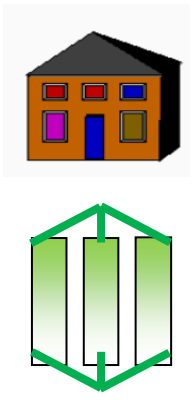
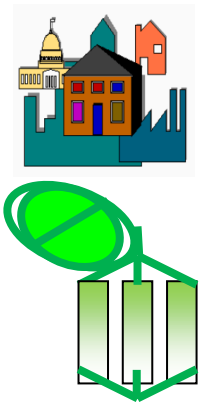
2.1.5.1.5. Soyutlanmış Yapı

Soyutlanmış yapı düzeyinde, bir önceki basamakta anlamlı bir bütün oluşturacak şekilde bir araya getirilen parçaların, daha yüksek bir soyutlama seviyesinde yeniden yapılandırılması söz konusudur (Leung, 2000). Bu düzeyde mevcut bilgilerin ötesinde, farklı bir konuya ya da yeni bir alana genellemeler yapılabilmektedir. Çalışılan konu ya da problem üst biliş düzeyinde anlaşılmakta ve farklı bir alana transfer edilebilmektedir (Thompson, 2007). Öğrenciler bu düzeyde, alternatif yaklaşımların farkına varır, kanıtlanmış kaynaklara atıf yaparak teoriler ve hipotezler üretir, genellemelere ulaşmak için tümdengelim akıl yürütme yollarını ve kombinasyonel düşünme süreçlerini kullanırlar (Lake, 1999). Kuram oluşturmak, genellemeler yapmak, tahmin etmek, hipotez kurmak, değerlendirmek, yansıtmak, teoriyi yeni bir alana uygulamak, derinlemesine incelemek soyutlanmış yapı düzeyinin gösterge fiillerinden bazılarıdır (Brabrand, 2008; Brabrand ve Dahl, 2009).

SOLO taksonomisini oluşturan düzeylerin tanımlanmasında, öğrencilerin sorulara vermiş oldukları cevapların yapısal karmaşıklığı esas alınmaktadır (Jimoyiannis, 2011). SOLO hiyerarşisinde üst düzeylere doğru ilerledikçe tutarlılık, ilişkilendirme ve çok yönlü düşünme artmaktadır (Biggs ve Collis, 1982; Hattie ve Purdie, 1998). Yapı öncesinden tek yönlü yapı düzeyine geçerken ilk olarak üzerinde çalışılan konunun tek bir yönü dikkate alınmaktadır. Tek yönlü yapı düzeyinden çok

yönlü yapı düzeyine geçişte ise konunun birden fazla yönüne odaklanılmaktadır (Brabrand ve Dahl, 2009). Bu nedenle; yapı öncesi, tek yönlü yapı ve çok yönlü yapı düzeylerinde, öğrencinin konuya ilişkin öğrenmelerinde niceliksel bir artış söz konusudur (Goel, 2011). Bu üç düzeyde gerçekleşen öğrenmeler yüzeysel öğrenmelerdir (Van Rossum ve Schenk, 1984). Yüzeysel öğrenmelerde, öğrenci olayların altında yatan ilkeler ile ilgilenmeden; yalnızca ezberlemeye odaklanmakta (Biggs, 1979) ve konuları tutarlı bir bütün olarak değil; birbirinden bağımsız parçalar olarak görmektedir (Scouller, 1998). Çok yönlü yapı düzeyinden ilişkisel yapı düzeyine geçerken konunun farklı yönleri arasında ilişkiler kurulur. Çok yönlü yapı düzeyinde listelenen özellikler, ilişkisel yapı düzeyinde anlamlı bir yapı oluşturmak üzere bir araya getirilir. İlişkisel yapı düzeyinden soyutlanmış yapı düzeyine geçerken, öğrenci mevcut bilgilerin ötesinde akıl yürütülebilir ve genellemelere ulaşılabilmektedir (Jurdak, 1991). Dolayısıyla, ilişkisel yapı ve soyutlanmış yapı düzeylerinde, öğrencinin öğrenmelerinde niteliksel bir artış meydana gelmektedir (Brabrand ve Dahl, 2009; Wadhwa, 2008). Öğrenme çıktılarında niteliksel bir dönüşümün olduğu bu düzeylerde, derin öğrenmeler gerçekleşmektedir (Van Rossum ve Schenk, 1984). Derin öğrenmelerde öğrenciler, karşılaştıkları yeni bilgileri, kişisel tecrübeleri ve eski bilgileri ile ilişkilendirerek anlamlandırmaya çalışmaktadır (Offir, Lev ve Bezalel, 2008). SOLO taksonomisini oluşturan düzeyler, bu düzeylerin temel özellikleri, her bir düzeyin gösterge fiilleri ve şekilsel gösterimleri Tablo 2.6'da sunulmuştur.

Tablo 2.6. SOLO taksonomisinin düzeyleri, bu düzeylerin gösterge fiilleri ve şekilsel gösterimleri

	Niceliksel Artış ve Yüzeysel Öğrenme			Niteliksel Artış ve Derin Öğrenme	
SOLO Düzeyleri	Yapı Öncesi	Tek Yönlü Yapı	Çok Yönlü Yapı	İlişkisel Yapı	Soyutlanmış Yapı
Temel Özellikleri	Üzerinde çalışılan konu ile ilgili öğrenilenler yanlıştır ya da herhangi bir şey öğrenilmemiştir.	Üzerinde çalışılan konunun tek bir yönüne odaklanılır.	Üzerinde çalışılan konunun iki ya da daha fazla yönü anlaşılır, fakat parçalar arasında ilişki kurulamaz.	Üzerinde çalışılan konunun farklı yönleri birbirleri ile ilişkilendirilir, bu sayede tutarlı bir yapıya sahip bir bütün elde edilmektedir.	Mevcut bilgilerin ötesinde akıl yürütülebilir ve genellemelere ulaşılabilir. Farklı bir alana transfer edebilme söz konusudur.
Gösterge Fiilleri		-Açıklamak -Tanımlamak -Ezberlemek -Basit bir işlemi uygulamak -Adlandırmak -Sıralamak -Saymak	-Birleştirmek -Sınıflandırmak -Numaralandırmak -Listelemek -Tanımlamak -Metaforik konuşmak -Planlamak -Algoritmaları ve yöntemleri uygulamak	-Analiz etmek -Karşılaştırmak -Birleştirmek -İlişkilendirmek -Sebep ve sonuçları açıklamak -Verilen bir teoriyi ilgili alana uygulamak	-Kuram oluşturmak -Genellemeler yapmak -Tahmin etmek -Hipotez kurmak -Değerlendirmek -Yansıtmak -Teoriyi yeni bir alana uygulamak -Tartışmak -Derinlemesine incelemek
Şekilsel Gösterimi					

SOLO taksonomisi Biggs ve Collis (1982) tarafından beş düzeyli olarak tanımlanmasına rağmen, daha sonra yapılan çalışmalarda, düzeylerin yeniden yapılandırılmasına ilişkin farklı araştırmacılar tarafından değişik öneriler getirilmiştir. Burnett (1999), çok yönlü yapı düzeyini düşük, orta ve yüksek şeklinde üç kategoriye, ilişkisel yapı düzeyini ise düşük ve yüksek şeklinde iki kategoriye ayırarak SOLO taksonomisini sekiz düzeyli bir yapı olarak yeniden tanımlamıştır. Bu tanımlamada yapı öncesi, tek yönlü yapı ve soyutlanmış yapı düzeylerinin özellikleri, beş düzeyden oluşan orijinal taksonomi ile aynıdır. Çok yönlü yapı düzeyi, üzerinde çalışılan konu hakkında dikkate alınan durumların sayısı ve bu durumların ne derece

ayrıntılı açıklandığına bağlı olarak üç düzeye ayrılmıştır. Çok yönlü yapı-düşük düzeyinde, sorulan soru ile ilgili iki ya da üç özelik birbirinden bağımsız olarak ele alınmaktadır. Ancak dikkate alınan özellikler hakkında bir ilişkilendirme ve açıklama yoktur. Çok yönlü yapı-orta düzeyinde sorulan soruya ilişkin çok sayıda durum dikkate alınmaktadır. Ancak bu düzeyde de ele alınan durumlara ilişkin herhangi bir ilişkilendirme ve açıklama bulunmamaktadır. Çok yönlü yapı-yüksek düzeyinde ise, sorulan soruya ilişkin birçok özellik göz önünde bulundurulmakta ve ele alınan durumlar örneklerle ayrıntılı bir biçimde açıklanmaktadır. Ayrıca, çok yönlü yapı-yüksek düzeyinde, öğrenci ele aldığı durumları birbiri ile ilişkilendirmeye yönelik bir çaba göstermektedir. İlişkisel yapı düzeyinin iki ayrı düzeye ayrılmasında ise, yapılan ilişkilendirmelerin ve genellemelerin derecesi esas alınmıştır. İlişkisel yapı-düşük düzeyinde, üzerinde çalışılan konunun bazı yönleri birbirleri ile ilişkilendirilmekte ve genellemeler birbiri ile ilişkilendirilen durumlar üzerinden yapılmaktadır. Ancak, konunun ya da problemin bazı yönleri, ilişkilendirmenin dışında kalmakta ve bütüne dâhil edilmemektedir. Diğer bir ifadeyle, ilişkisel yapı-düşük düzeyinde öğrenilen durumlar, çoğunlukla anlamlı bir kavram oluşturmak üzere bir araya getirilir. Bununla birlikte, bütünden sapan ve yapının geneli içine alınmayan birkaç parça bulunmaktadır. İlişkisel yapı yüksek düzeyinde, öğrenilen noktalar anlamlı bir kavram oluşturmak üzere bir araya getirilir. Öğrenci, soruya verdiği cevabı belirgin ve sağlam kanıtlarla desteklemektedir. Bu düzeyde, anlamlı bir bütün oluşturulmakta ve bütünün dışında kalan herhangi bir parça bulunmamaktadır.

Burnett'e benzer şekilde, Chan vd. (2002) de çok yönlü yapı ve ilişkisel yapı düzeylerinin alt kategorilere ayrılmasını önermiştir. Tıpkı Burnett gibi Chan vd. (2002) de, çok yönlü yapı düzeyini düşük, orta ve yüksek şeklinde üç düzeye ayırmıştır. Ancak ilişkisel yapı düzeyini iki kategoriye ayıran Burnett'ten (1999) farklı olarak, Chan vd. (2002) bu düzeyi de düşük, orta ve yüksek şeklinde üç düzeye ayırarak incelemiştir. Dolayısıyla Chan vd. (2002), SOLO taksonomisinin yapı öncesi, tek yönlü yapı, çok yönlü yapı-düşük, çok yönlü yapı-orta, çok yönlü yapı-yüksek, ilişkisel yapı-düşük, ilişkisel yapı-orta, ilişkisel yapı-yüksek ve soyutlanmış yapı olmak üzere dokuz düzeyli olarak tanımlamıştır. Dokuz düzeyli bu yapıda ilişkisel yapı düzeyleri dışındaki tüm düzeyler Burnett (1999) tarafından önerilen sekiz düzeyli yapı ile aynı özeliklere sahiptir. İlişkisel yapı düzeyi ise, Burnett'in (1999) yaptığı tanımlamaya göre daha detaylı olarak açıklanmıştır. Chan vd.'ne

(2002) göre, ilişkisel yapı-düşük düzeyinde, üzerinde çalışılan konunun iki ya üç yönü birbiri ile ilişkilendirilmekte ve genellemeler yalnızca bu iki ya da üç durumdan hareketle yapılmaktadır. İlişkisel yapı-orta düzeyinde, ilişkilendirmeler ve genellemeler yapılırken üzerinde çalışılan konunun birçok farklı özelliği göz önünde bulundurulmaktadır. Ancak yine de ilişkilendirme ve genellemelere dâhil edilmeyen durumlar bulunmaktadır. İlişkisel yapı-yüksek düzeyinde ise, ilişkilendirmeler ve genellemeler üzerinde çalışılan konunun bütün yönleri dikkate alınarak yapılmaktadır.

2.1.5.2. SOLO Taksonomisinin Avantajları Nelerdir?

Birçok eğitim sisteminde değerlendirme işlemi niceliksel olarak yapılmakta ve öğrencinin ne kadar öğrendiğine odaklanılmaktadır. Yalnızca öğrenme miktarına odaklanan değerlendirme anlayışları nicel değerlendirmeler olarak ifade edilmektedir. Nicel değerlendirme anlayışları, öğrenmenin derinliği belirlemede yetersiz kalmaktadır. Dolayısıyla, öğrencilerin ne kadar öğrendiklerinin yanı sıra, bu öğrenmelerin niteliğini de belirleyebilecek bir model ihtiyaç haline gelmektedir (Bhattacharyya, Bhattacharya ve Mitra, 2012). SOLO taksonomisi bu ihtiyacı karşılayacak bir model olarak ileri sürülmüştür (Slack vd., 2003). SOLO taksonomisine göre, öğrenmenin hem nicel (yüzeysel öğrenme), hem de nitel yönü (derin öğrenme) bulunmaktadır (Maddrell, 2011). Öğrencinin herhangi bir soruya verdiği cevaptaki niceliksel ve niteliksel özellikler bir araya gelerek cevabın yapısal karmaşıklığını oluşturmaktadır. Öğrencinin verdiği cevaptaki ayrıntının miktarı yapısal karmaşıklığın niceliksel yönünü yansıtmaktadır. Öğrencinin cevabında listelediği ayrıntıları birbirileri ile ne kadar iyi ilişkilendirdiği ise, yapısal karmaşıklığın niteliksel yönünü meydana getirmektedir (Biggs ve Tang, 2007). SOLO taksonomisi kullanılarak öğrencinin verdiği cevaplar, yapısal karmaşıklığın hem nicel hem de nitel yönü açısından değerlendirilebilmektedir (Leung, 2000).

SOLO taksonomisinin ikinci bir avantajı, cevabın doğru ya da yanlışlığından ziyade, öğrencinin problemi nasıl anladığına ve problemin çözümünde nasıl bir yol izlediğine önem veren bilişsel bir model olmasıdır (Jimoyiannis, 2011). Bir başka ifadeyle, SOLO taksonomisi *öğrenci konuyu anlamıştır* ya da *öğrenci konuyu anlamamıştır* şeklinde bir değerlendirme yapmak yerine; konunun ne düzeyde anlaşıldığı ile ilgilenmektedir (Ireson, 2008). Buna göre, SOLO taksonomisi kullanılarak yapılan değerlendirmelerin, "*öğrenci konuyu ne düzeyde anlamıştır"*

sorusuna yanıt olduğu söylenebilir. SOLO taksonomisinin bu özelliği, öğrencilerin öğrenme sürecindeki eksikliklerini görmeye yardımcı olmakta ve değerlendirme sürecinde kısmi puanlamalar yapmaya imkân tanımaktadır. SOLO taksonomisi kullanılarak, öğrencilerin öğrenme eksikliklerinin belirlenebilmesi, SOLO taksonomisini düzey belirlemeye yönelik değerlendirmelerin yanı sıra biçimlendirici değerlendirmeler için de uygun bir araç haline getirmektedir (Hattie ve Purdie, 1998).

SOLO taksonomisinin avantajlarından biri de, ilköğretimden yükseköğretime kadar birçok farklı eğitim kademesinde kullanılabilmesidir. Örneğin; Akkaş (2009), ve Bağdat (2013) ilköğretim öğrencileriyle, Tuna (2011) ve Jimoyiannis (2011) lise öğrencileriyle, Groth ve Bergner (2006) ile Çelik (2007) lisans öğrencileriyle ve Chan vd. (2002) lisansüstü düzeydeki öğrencilerle çalışmış ve yaptıkları analizlerde SOLO taksonomisini kullanmışlardır. SOLO taksonomisi farklı eğitim kademelerinde kullanılabildiği gibi, birçok farklı derste de kullanılabilmektedir (Kanuka, 2005, 2011). SOLO taksonomisinin farklı derslerde kullanımını örneklendirmek gerekirse, Biggs ve Collis (1982) tarafından yapılan araştırmada tarih, Courtney (1986) tarafından yapılan araştırmada coğrafya, Holbrook (1989) tarafından yapılan çalışmada kimya, Collis ve Romberg (1992) ile Lian ve Idris (2006) tarafından yapılan çalışmalarda matematik, Lake (1999) tarafından yapılan araştırmada biyoloji, Prosser, Trigwell ve Waterhouse (2000) tarafından yapılan araştırmada fizik, Chan vd. (2002) tarafından yapılan çalışma kompozisyon yazma, Serow (2007) tarafından yapılan araştırmada ise geometri dersi öğrenme çıktılarının değerlendirilmesinde SOLO taksonomisinden yararlanılmıştır.

SOLO taksonomisinin avantajlarından bir diğeri, ölçme araçlarında yer alacak maddelerin yazımında kullanılabilmesidir. Başarı testleri geliştirilirken, testte yer alacak maddeler SOLO taksonomisinin düzeylerine uygun olarak hazırlanabilmektedir. Bununla birlikte, SOLO taksonomisi sadece bir madde yazma yöntemi değildir. Maddelerin puanlanmasında da SOLO taksonomisinden yararlanılabilmektedir. SOLO taksonomisi kullanılarak öğrencinin verdiği cevaptaki fikirlerin sayısı (1=tek yönlü yapı, 2=çok yönlü yapı) ya da cevapta kullanılan ilişkilendirmenin derecesi (3=doğrudan ilişkilendirmeler ya da 4=farklı bir alana yapılan genellemeler) belirlenebilir. Daha sonra, verilen cevabın SOLO taksonomisinin hangi düzeyine denk geldiği tespit edilip, bu doğrultuda puanlamalar yapılabilmektedir. SOLO taksonomisinin avantajları maddelerin puanlanması ile de

sınırlı kalmamaktadır. SOLO taksonomisi, madde yazımı ve maddelerin puanlanmasına ek olarak, ölçme ve değerlendirme ile ilgili diğer pek çok önemli konuya da çözüm olabilmektedir (Hattie ve Purdie, 1998). Örneğin, bilgisayar ortamında bireye uyarlanmış testlerin uygulanmasında SOLO taksonomisinden yararlanılabilmektedir. Bireye uyarlanmış testlerde, madde bankasındaki maddeler SOLO taksonomisinin düzeylerine göre sınıflandırılabilir. Daha sonra, bireye sorulacak maddelerin ne zaman sona erdirileceğini belirleyen durdurma kuralına "*SOLO taksonomisinin tüm düzeyleri ile ilgili sorular sorulduktan sonra testin sona erdirilmesi*" şeklinde bir şart eklenebilir. Bu sayede, bilgisayar ortamında bireye uyarlanmış testler öğrencilerin hem nicel hem de nitel öğrenmelerini değerlendirebilecek şekilde uygulanabilmektedir (Hattie ve Purdie, 1998).

SOLO taksonomisinin ölçme değerlendirme süreçleri açısından sağladığı bir diğer avantaj, öğrencilere kendi çalışmalarını analiz etme ve ilerlemelerini görme imkânı sunmasıdır. SOLO taksonomisinde, değerlendirmede kullanılacak ölçütler açık bir biçimde tanımlanmaktadır (Killen, 2009). Öğrenciler SOLO taksonomisinde belirtilen bu ölçütleri kullanarak öz değerlendirme yapabilmekte ve öğrenme sürecindeki gelişimleri hakkında fikir edinebilmektedir. Öğrencinin öğrenme sürecindeki gelişimini görmesi, daha sonraki öğrenmelere yönelik önemli bir motivasyon kaynağı olmaktadır (Leung, 2000). Hattie ve Purdie'ye (1994) göre; SOLO taksonomisinin ölçme değerlendirme sürecine yönelik olarak sağladığı avantajlardan biri de, diğer taksonomilere göre daha yüksek puanlayıcı güvenirliğine sahip olmasıdır. Hattie ve Purdie (1994) tarafından yapılan çalışmada 30 öğretmen, bir başarı testini oluşturan 19 maddeyi Bloom ve SOLO taksonomisine göre düzeylere ayırmıştır. Öğretmenlerin yarısı, maddeleri Bloom taksonomisine göre düzeylere ayırırken, diğer yarısı maddelerin düzeylere ayrılmasında SOLO taksonomisini esas almışlardır. SOLO taksonomisine göre yapılan sınıflamada öğretmenler arasında %60 oranında mutlak uyum belirlenmiştir. Öğretmenler arasında, en fazla bir düzey fark bulunan sınıflamaların oranı ise %96 olarak saptanmıştır. Diğer taraftan, Bloom taksonomisine göre yapılan sınıflamada, öğretmenler arasındaki mutlak uyum oranı %40 olarak bulunmuştur. Öğretmenler arasında en fazla bir düzey fark bulunan sınıflamaların oranı ise %75 olarak tespit edilmiştir. Hundzynski (2008) tarafından yapılan araştırmada, SOLO taksonomisi kullanılarak gerçekleştirilen değerlendirmelerde puanlayıcılar arası güvenirlik incelenmiştir. Puanlayıcılar arası güvenirlik katsayısını hesaplamak için basit uyum

yüzdesinden yararlanılmıştır. Araştırmada, SOLO taksonomisi kullanılarak yapılan değerlendirme işlemine yönelik puanlayıcı güvenirliği .87 olarak bulunmuştur. Chan vd. (2002) tarafından yapılan araştırmada ise, SOLO ve Bloom taksonomisi kullanılarak yapılan değerlendirmelerden hangisinin daha ayırt edici sonuçlar verdiği incelenmiştir. Araştırmada yetenek düzeyleri farklı olan öğrencileri birbirinden ayırt etmede, SOLO taksonomisinin Bloom taksonomisine göre daha etkili bir araç olduğu belirlenmiştir. Chan vd. (2002), bu sonuca dayanarak, SOLO taksonomisini Bloom taksonomisine göre daha kullanıcı dostu bir model olarak nitelendirmiştir.

SOLO taksonomisi, ölçme değerlendirme sürecine ilişkin sunduğu avantajların yanı sıra, öğretimin planlanması ve bir dersin hedeflerinin bilişsel düzeylere ayrılması gibi birçok farklı amaçla da kullanılabilmektedir (Hattie ve Purdie, 1998). Söz gelimi, Alsaadi (2011) tarafından yapılan çalışmada, İngiltere ve Katar'da uygulanmakta olan matematik öğretim programlarındaki kazanımlar SOLO taksonomisinin düzeyleri referans alınarak analiz edilmiştir. Peter ve Alberto'nun (2013) yaptığı araştırmada, Avusturalya'da uygulanmakta olan Kimya Dersi Öğretim Programı (12. sınıf) kazanımları ile sınav soruları SOLO taksonomisinin düzeylerine göre incelenmiştir. Gezer ve İlhan (2014) tarafından yapılan araştırmada, Vatandaşlık Dersi Öğretim Programı'nda yer alan kazanımlar ile ders kitabı değerlendirme sorularının SOLO taksonomisinin düzeylerine göre analizi yapılmıştır. Gezer, İlhan, Öner Sünkür ve Çetin'in (2014) yaptığı araştırmada ise sosyal bilgiler dersi sınav soruları SOLO taksonomisinin düzeylerine göre incelenmiştir.

SOLO taksonomisi öğrencilerin nasıl öğrendiğine ve öğretmenlerin öğretim sürecinde nelere dikkat etmeleri gerektiğine ilişkin bir rehber sunmaktadır. Bu sayede, öğrencilerin SOLO taksonomisinin bir düzeyinden bir üst düzeye ilerlemesi için uygun öğrenme materyallerin seçilmesi ve öğretim durumlarını hazırlanması konusunda öğretmenlere destek olmaktadır. Örneğin, bir konu ile ilgili olarak bir öğrencinin tek yönlü yapı düzeyinde bulunduğu belirlenmiş olsun. Bu durumda, öğretmen üst düzeyler için karşılanması gereken standartlardan yararlanarak öğrenciyi ilişkisel yapı ya da soyutlanmış yapı düzeylerine taşımaya yardımcı olacak şekilde sınıf aktivitelerini ve dersleri planlayabilmektedir. Bu özelliğinden dolayı, SOLO taksonomisi öğrenme süreci ile değerlendirme sürecini birleştiren bir model olarak ifade edilmektedir (Hattie ve Purdie, 1998).

2.1.5.3. Sınav Sorularının SOLO Taksonomisine Göre Hazırlanması ve Puanlanması

SOLO taksonomisinin en önemli avantajlarından biri test maddelerinin hazırlanmasında kolaylıkla kullanılabilen bir model olmasıdır. Maddeleri SOLO taksonomisine göre yapılandırmanın üç yolu bulunmaktadır. Sorular belirli bir SOLO düzeyini ortaya çıkaracak şekilde düzenlenebilir, cevaplar SOLO düzeylerine bağlı olarak puanlanabilir ya da bu iki yöntem bir arada kullanılabilir (Hattie ve Purdie, 1998). Aşağıda, SOLO taksonomisinin farklı düzeylerine karşılık gelen açık uçlu soru örnekleri sunulmuştur. Örnek olarak sunulan sorular MEB tarafından kullanılan ders kitaplarından (Aşan, 2014; MEB, 2014a; Ülker, 2014; Ünver, 2014) alınmıştır.

Tablo 2.7. SOLO taksonomisinin farklı düzeylerine karşılık gelen açık uçlu madde örnekleri

Ders	Sınıf	Değerlendirme Sorusu	SOLO Taksonomisi Düzeyi
Türkiye Cumhuriyeti Inkilap Tarihi ve Atatürkçülük	8. Sınıf	I. Dünya Savaşı başladığında Osmanlı Mebusan Meclisi'nde vekil olarak görev yapmaktasınız. Savaşa katılan devletlerin Osmanlı Devleti ile ilgili tarihi hayallerini ve planlarını da dikkate aldığınızda, I. Dünya Savaşı'na girme konusunda neler yapardınız? Neden?	Soyutlanmış Yapı
Vatandaşlık ve Demokrasi Eğitimi	8. Sınıf	Hak ve özgürlüklerinize birer örnek vererek bunları sorumluluklarınızla ilişkilendiriniz.	İlişkisel Yapı
Fen ve Teknoloji	7. Sınıf	Şekerli su çözeltisi deriştirmek için neler yapabiliriz açıklayınız.	Çok Yönlü Yapı
Fen ve Teknoloji	8. Sınıf	Canlıların DNA yapılarındaki değişim ……... olarak adlandırılır.	Tek Yönlü Yapı

- Tablo 2.7'deki değerlendirme sorularından ilki; SOLO taksonomisinin *tartışmak, tahmin etmek* ve *varsayımda bulunmak* gösterge fiillerini yansıtmaktadır. Dolayısıyla, öğrencinin mevcut bilgilerin ötesinde fikir üretmesini gerektiren bu soru soyutlanmış yapı düzeyine karşılık gelmektedir.
- Tablodaki ikinci değerlendirme sorusunun cevaplandırılabilmesi için öğrencinin hak, özgürlük ve sorumluluk kavramlarını bilmesi ve bu kavramları birbiriyle ilişkilendirmesi gerekmektedir. Buna göre, İlişkilendirmek gösterge fiili ile eşleşen bu madde SOLO taksonomisinin ilişkisel yapı düzeyine denk gelmektedir.
- Üçüncü değerlendirme sorusunda öğrencinin konu ile ilgili birden fazla yönü bilmesi ve bunları listelemesi beklenmektedir. Dolayısıyla, bu değerlendirme

sorusu çok yönlü yapı düzeyinin listelemek gösterge fiili ile örtüşmekte ve çok yönlü yapı düzeyine karşılık gelmektedir.

- Öğrencinin konu ile ilgili tek bir bilgiyi hatırlaması tablodaki dördüncü değerlendirme sorusunu cevaplayabilmesi için yeterlidir. Buna göre, *ezberlemek* gösterge fiilleri ile eşleşen bu soru tek yönlü yapı düzeyinde yer almaktadır.

SOLO taksonomisi açık uçlu soruların yapılandırılmasında kullanılabildiği gibi, çoktan seçmeli sorular da SOLO taksonomisine uygun olarak hazırlanabilmektedir. Aşağıda, MEB tarafından kullanılan ders kitaplarından (Aşan, 2014; MEB, 2014b; Ülker, 2014) alınan ve SOLO taksonomisinin farklı düzeylerine karşılık gelen çoktan seçmeli soru örnekleri sunulmuştur.

Tablo 2.8. SOLO taksonomisinin farklı düzeylerine karşılık gelen çoktan seçmeli soru örnekleri

Ders	Sınıf	Değerlendirme Sorusu	SOLO Taksonomisi Düzeyi
TC İnkılap Tarihi ve Atatürkçülük	8. Sınıf	Aşağıdakilerden hangisi, Türkiye'nin Avrupa Ekonomi Topluluğu (AET) ile Ankara Antlaşmasını imzaladığı tarihtir? A) 12 Haziran 1959 B) 25 Haziran 1963 C) 12 Eylül 1963 D) 6 Mart 1995	Tek Yönlü Yapı
Vatandaşlık ve Demokrasi Eğitimi	8. Sınıf	Hukukun üstünlüğünde toplum; devletin etkisinde olmadığından özgür, çoğulcu ve katılımcıdır, egemenliğin hem sahibi hem de tek kullanıcısıdır. **Yukarıdaki ifadeden yola çıkarak hukukun üstünlüğünün egemen olduğu bir ülkede aşağıdaki yargılardan hangisi söz konusu olamaz?** A) İnsan hak ve özgürlükleri geliştirilebilir. B) Siyasal özgürlüğe dayanan demokratik bir rejimin varlığı sağlanmıştır. C) Egemenlik bütünüyle devlete aittir, devredilemez. D) Bireyler özgür, girişimci ve katılımcıdır.	İlişkisel Yapı
Sosyal Bilgiler	6. Sınıf	Aşağıdaki diyaloğu okuyunuz - **Burcu:** Bilimsel araştırma yapmak için öncelikle konuyu belirlemeliyiz. - **Emre:** Sorunun çözümüne yönelik varsayımlar yazmalıyız. - **Burcu:** Konuyu çeşitli kaynaklardan araştırmalıyız. - **Emre:** Araştırmada elde edilen verileri rapor haline getirmeliyiz. Bu konuşmada hangi bilimsel araştırma basamağından söz edilmemiştir? A) Varsayımların test edilmesi B) Varsayımlarda bulunulması C) Araştırma konusunun belirlenmesi D) Araştırma raporunun yazılması	Çok Yönlü Yapı

Tablo 2.8'de yer alan çoktan seçmeli soruların her biri farklı bir SOLO düzeyini ortaya çıkaracak şekilde yapılandırılmıştır. Bazı durumlarda ise, soru aynı olmasına rağmen, öğrencilerin verdikleri cevaplar SOLO taksonomisinin farklı düzeylerine karşılık gelebilmektedir. Aşağıda tüm öğrencilerin aynı soruyu cevaplandıkları, ancak öğrencilerin sorulara verdikleri cevapların SOLO taksonomisinin farklı düzeylerinde yer aldığı bir örnek sunulmuştur. Bu örnek Pegg ve Coddy'nin (1993) çalışmasından alınmıştır (Akt: Çelik, 2007).

Tablo 2.9. SOLO taksonomisi kullanılarak puanlanan açık uçlu soru örneği

Soru: p bir reel sayı olmak üzere $1/p > p$ ifadesinin ne anlama geldiğini tartışınız	Açıklama	Düzey
p bir kesir olmalıdır. Çünkü $p=\frac{1}{2}$ için $\frac{1}{\frac{1}{2}} > 1$ ⇨ 2>1	Her iki cevapta da tek bir durum üzerine odaklanılmıştır. Birinci cevap özellikle 1'den küçük pozitif reel sayılar üzerine odaklanmıştır. Birinci cevaptaki "kesir" kelimesi ile 1'den küçük kesirler kastedilmiştir, $p\neq0$ durumu irdelenmemiştir. İkinci cevapta yalnızca sayının pozitif ya da negatif olmasına odaklanılmıştır. Tek yönlü yapı düzeyinde öğrenci cevabını sayısal örneklerden yararlanarak açıklama ihtiyacı hissetmektedir.	Tek Yönlü Yapı
p pozitif bir sayı olduğunda, denklem doğru değildir. *p* negatif bir sayı olduğunda ise, denklem doğru olmaktadır.		
$p\neq0$, $p<1$ ve $p>0$ olmalıdır. Örneğin, $\frac{1}{2}$ için ifade doğru olmaktadır.	Yandaki cevapların her birinde problemle ilişkili iki veya daha fazla durum düşünülmüştür. Bununla birlikte cevaplarda açık bir tutarlılık bulunmamaktadır. Dolayısıyla yandaki her üç cevap da çok yönlü yapı düzeyine karşılık gelmektedir. Birinci cevapta, pozitif reel sayılar düşünülmüş, ancak *p*'nin alabileceği değerlerin hangi aralıkta bulunduğu tam olarak ifade edilmemiştir. İkinci cevapta negatif reel sayılar üzerine odaklanılmıştır. Üçüncü cevapta, hem kesirler hem de negatif sayılara odaklanılmıştır. Ancak bunlar birbirleri ile ilişkilendirilip *p*'nin alabileceği değer aralığı tam olarak tanımlanamamıştır. Çok yönlü yapı seviyesinde de tek yönlü yapı düzeyinde olduğu gibi öğrenci somut örneklerden yararlanma ihtiyacı hissedebilmektedir.	Çok Yönlü Yapı
p sıfıra eşit olmamalıdır. $p>0$, $p\neq-1$ olduğunda yanlıştır. $p<0$ olduğunda ifade doğru olmaktadır. Ayrıca $p<-1$ şartını sağlayan tüm değerler için ifade doğrudur.		
p uygun bir kesirli sayı olduğunda ifade doğru olacaktır. *p*'nin tüm negatif değerleri için ifade doğru olmaktadır. Sıfır dışında tüm $p<0$ değerleri için ifade doğrudur.		
$0<p<1$ ve $p< -1$	Öğrenci problemde ne istendiğini tam olarak anlamıştır. Problem ile ilgili farklı durumları belirleyip bunları birbiriyle ilişkilendirilebilmektedir. Çözümünün gösteriminde sembollerden yararlanmaktadır. Çözümünü açıklamak için sayısal değerlerden yararlanma ihtiyacı hissetmemektedir.	İlişkisel Yapı
$p\neq0$, $\{p: p< -1\} \cup \{p: 0<p<1\}$		

Örneklerde görüldüğü gibi, SOLO taksonomisi hem çoktan seçmeli hem de açık uçlu soruların hazırlamasında kullanılabilmektedir. Bununla birlikte, soruları yanıtlarken öğrencilere daha fazla özgürlük tanıyan açık uçlu soruların SOLO taksonomisine daha uygun olduğu ifade edilmektedir (Leung, 2000). Çünkü çoktan seçmeli sorularda, herhangi bir maddeye aynı yanıtı veren iki öğrencinin cevaba ulaşmak için izledikleri yollar farklı olabilmektedir. Diğer bir ifadeyle, iki öğrenci SOLO taksonomisinin farklı düzeylerinde bulunmasına ve çoktan seçmeli bir sorunun çözümü için değişik yollar izlemesine rağmen, soruya verdikleri cevap aynı olabilmektedir. Dolayısıyla çoktan seçmeli sorular açık uçlu sorulara göre, SOLO taksonomisinin farklı düzeylerinde bulunan öğrencileri birbirinden ayırt etmede yetersiz kalabilmektedir. Aşağıda, bu durumu örneklendiren çoktan seçmeli bir soru yer almaktadır.

Tablo 2.10. Problem çözme sürecinde takip edilen işlem basamaklarına göre cevabın karşılık geldiği SOLO düzeyi

3(X + 2) = 34 – X eşitliğinde yer alan X'in değeri nedir? A) 4 B) 5 C) 6 D) 7
Bu soruyu doğru yanıtlayarak 7 cevabını veren iki öğrencinin doğru cevaba ulaşmak için izledikleri yollar farklı olabilir. Örneğin, bir öğrenci seçeneklerde verilen değerleri sunulan eşitlikte yerine koyarak doğru cevaba ulaşmış olabilir. Cevaba bu şekilde ulaşan öğrencinin bulunduğu düzey çok yönlü yapı olarak belirlenir. Bir başka öğrenci verilen eşitliği, 3X+6 = 34–X ve X = $\frac{34-6}{4}$ şeklinde yazıp bilinmeyeni tek başına bırakarak doğru cevaba ulaşabilir. Doğru cevaba bu şekilde ulaşan öğrencinin bulunduğu düzey ise, ilişkisel yapı olarak tespit edilir. Görüldüğü gibi, iki öğrenci SOLO taksonomisinin farklı düzeylerinde bulunmasına ve sorunun çözümüne ulaşmak için farklı yollar izlemesine rağmen, aynı sonuca ulaşabilmektedir. Dolayısıyla öğrencinin sorunun cevabına ulaşmak için nasıl bir yol izlediğini görmeye olanak tanıyan açık uçlu soruların çoktan seçmeli sorulara kıyasla SOLO taksonomisine daha uygun olduğu söylenebilir.

Yukarıdaki örnekten de anlaşılacağı üzere, çoktan seçmeli sorular ile SOLO taksonomisinin farklı düzeylerinde bulunan öğrencilerin birbirinden başarılı bir biçimde ayırt edilmesi oldukça zordur. Dolayısıyla, öğrencinin herhangi bir soruya verdiği cevabın, SOLO taksonomisinin hangi düzeyine karşılık geldiğinin tam olarak belirlenebilmesi için açık uçlu sorulardan yararlanılmalıdır (Leung, 2000). Öğrencinin açık uçlu bir soruya verdiği cevap ile bu cevabın karşılık geldiği SOLO düzeyinin eşleştirilmesi puanlayıcı kararlarına göre yapılmaktadır. Bu durum, SOLO taksonomisine dayalı rubrikler ile puanlanan açık uçlu sorularda puanlayıcı güvenirliğinin belirlenmesini gerekli kılmaktadır.

2.1.6. Puanlayıcı Güvenirliği

Ölçme ve değerlendirme çalışmalarının amacı ölçülen özelliğe ilişkin açık ve doğru kestirimler elde etmektir. Bununla birlikte, tüm ölçüm işlemlerinin mükemmel olması mümkün değildir. Ölçme işlemi ne kadar dikkatli yapılırsa tüm ölçmelerde bir miktar hatanın bulunması kaçınılmaz olmaktadır (Güler, 2012). Özellikle; doğa bilimleri, sağlık bilimleri ve sosyal bilimlerde elde edilen ölçümler nadiren kesin olarak kabul edilebilmektedir. Başarı, tutum, kaygı, zekâ ya da yönetsel yetenek gibi özelliklerin ölçülmesi sırasında, ölçme işlemine bir miktar hata karışmaktadır. Çünkü burada kestirilmeye çalışılan özelliğin kesin (mutlak) bir biçimde tanımlanması zordur ve bu özellikler çoğunlukla doğrudan gözlenemezler. Dolayısıyla, söz konusu özellikleri ölçtüğü kabul edilen ölçme araçları oluşturulmaktadır. Ancak çok sayıda faktör bu ölçüm sürecini etkilemekte ve elde edilen sonuçlarda hataya neden olan değişkenlikler üretmektedir (Cardinet, Johnson ve Pini, 2010). Ölçme işleminde hatalardan kaçınmak için hataların kaynakları ve bu hata kaynaklarının ölçme sonuçlarını nasıl etkilediği bilinmelidir (Turgut ve Baykul, 2012). Bu durum ölçme işleminde ve sonuçlarında hataya neden olan değişkenlik kaynaklarının araştırılmasını yani güvenirliğin test edilmesini gerekli kılmaktadır (Atılgan, 2004).

Güvenirlik ölçme araçlarından elde edilen puanlar arasındaki tutarlılık olarak tanımlanmaktadır (Moskal ve Leydens, 2000). Tutarlık kelimesine ilişkin işe vuruk (operasyonel) tanımları açısından gösterdikleri farklılığa bağlı olarak; test tekrar test güvenirliği, paralel (eş değer) form güvenirliği, iç tutarlık güvenirliği ve puanlayıcı güvenirliği gibi birçok farklı güvenirlik türü bulunmaktadır (Goodwin, 2001). Test tekrar test güvenirliği, bir ölçme aracının iki ya da daha çok uygulamasından elde edilen puanlar arasındaki tutarlığın bir ölçüsüdür. Bu tutarlılık ölçme aracının zamana karşı değişmezliğini göstermektedir. Dolayısıyla, test tekrar test güvenirliğinde, hata kaynağı iki uygulama arasındaki zaman olarak değerlendirilmektedir. Paralel form güvenirliği, aynı yapıyı ölçen iki eş değer formdan elde edilen puanlar arasındaki tutarlığın bir ölçüsüdür. Paralel form güvenirliğinde, hata kaynağı olarak ölçme aracı dikkate alınmaktadır (Atılgan, 2004). İç tutarlık güvenirliği, bir ölçekte yer alan maddeler arasındaki tutarlığın bir ölçüsüdür. İç tutarlılık güvenirliğinde, ölçme aracında yer alan maddeler hata kaynağı olarak değerlendirilmektedir (Güler, 2008). Puanlayıcı güvenirliği ile aynı bireye iki ya da daha fazla sayıda farklı puanlayıcı tarafından verilen ya da birey için aynı puanlayıcı tarafından farklı zamanlarda hesaplanan puanların tutarlılığı

kastedilmektedir (Moskal ve Leydens, 2000). Puanlayıcı güvenirliğinde hata kaynağı olarak puanlayıcılar dikkate alınmaktadır. Puanlama işleminin objektif olarak yapılamadığı ve değerlendirme işlemine puanlayıcıların öznel yargılarının karışabildiği durumlarda puanlayıcı güvenirliğinin hesaplanması gerekmektedir (Kan, 2009). Puanlayıcı güvenirliği, puanlayıcı içi güvenirlik (intra rater reliability) ve puanlayıcılar arası güvenirlik (inter rater reliability) olmak üzere iki başlıkta incelenmektedir (Jonsson ve Svingby, 2007).

Puanlayıcı içi güvenirlik, bir puanlayıcının kendi puanlamaları arasındaki tutarlığa yönelik bir ölçüdür (Barkaoui, 2008). Bu güvenirlik, bir puanlayıcının farklı zamanlarda aynı bireylerin kâğıtlarını birden fazla kez puanlaması ve elde edilen puanlar arasındaki tutarlığın ölçülmesiyle belirlenmektedir. Puanlayıcının yorgunluğu, kimi puanladığını bilmesi ve psikolojik durumundaki değişmeler gibi gelişigüzel (tesadüfi/random) faktörler puanlayıcının farklı zamanlarda yaptığı puanlamalar arasında uyumsuzluğun görülmesine sebep olabilmektedir (Güler, 2008). Puanlayıcılar arası güvenirlik ise, aynı performansa ilişkin iki ya da daha fazla sayıda farklı puanlayıcı tarafından verilen puanlar arasındaki tutarlığın bir ölçüsüdür (Wang, 2009). Puanlayıcılar arası güvenirlik analizleriyle, iki puanlayıcı arasındaki hata varyansı belirlenebilmektedir. Belirlenen bu hata varyansı gözlenen puanlardaki varyanstan çıkarılarak, gözlenen puanlardaki varyansın ne kadarının gerçek varyanstan kaynaklandığı ortaya konulabilir (Novick, 1996). Örneğin, puanlayıcılar arası güvenirliğin .80 olarak hesaplanması gözlenen varyansın %80'inin ölçülmek istenen özelliğe ait değişkenlikten; %20'sinin ise puanlayıcılar arası değişkenlikten kaynaklandığını göstermektedir (Hallgren, 2012). Buna göre, objektif olarak puanlanamayan ölçme araçları için uygulanan istatistiksel analizlerin ve bu istatistiksel analizlere ilişkin yorumların puanlayıcı hatalarından ciddi bir biçimde etkilendiği söylenebilir. Bu durumda, gözlenen varyanstaki değişkenliğin ne kadarının puanlayıcı hatasından kaynaklandığının belirlenmesi oldukça önemli bir konu haline gelmektedir. Dolayısıyla, puanlayıcı yargılarına dayalı olarak yapılan değerlendirmelerde, puanlayıcı güvenirliğinin hesaplanması gerekmektedir.

2.1.6.1. Puanlayıcı Güvenirliğinin Hesaplanması

Puanlayıcı güvenirliğinin belirlenmesinde kullanılan yöntemler, klasik test kuramına dayalı yöntemler ve madde tepki kuramına dayalı yöntemler olmak üzere iki başlıkta incelenmiştir (MacMillan, 2000).

2.1.6.1.1. Klasik test kuramına dayalı yöntemler

Klasik test kuramının temel varsayımı, bir ölçüme ilişkin gözlenen puanın, gerçek puan ile hata puanı bileşenlerinin toplamına eşit olmasıdır. Bu varsayım, X=T+E eşitliği ile gösterilmektedir. Bu eşitlik klasik test kuramının temel denklemi olarak ifade edilmektedir (Baykul, 2000). Bu eşitlikte, X bireyin testten aldığı puanı yani gözlenen puanı göstermektedir. T gerçek puanı temsil etmektedir. Gerçek puanlar, bir testin bireye sonsuz sayıda uygulanması (Kline, 2005; MacMillan, 2000) ve bu uygulamalar arasına öğrenme ya da başka faktörlerin etkisinin karışmaması durumunda, bireyin testten alacağı puanların aritmetik ortalamasıdır (MacMillan, 2000). Ancak, bir testin bir bireye sonsuz sayıda uygulanması mümkün olmadığından, gerçek puan hipotetik bir kavramdır (Kline, 2005). E ise, hata bileşenini temsil etmektedir. Puanlama işlemine; yorgunluk, kaygı, motivasyon eksikliği gibi bireysel faktörlerden; maddelerin açık olmaması, madde sayısının yetersiz olması gibi ölçme aracı ile ilgili faktörlerden ya da açık olmayan yönergeler, zamanın yeterli olmayışı gibi ölçme aracının uygulanması ile ilgili faktörlerden dolayı karışan hatalar, hata varyansını meydana getirmektedir (Goodwin, 2001). Objektif olarak puanlanamayan testlerde hata varyansına neden olan faktörlerden biri de puanlayıcılar ile ilgili faktörlerdir. Klasik test kuramında, puanlayıcı güvenirliğinin belirlenmesinde, basit uyum yüzdesi, Cohen'in kappası (Cohen, 1960), ağırlıklandırılmış kappa (Cohen, 1968), Fleiss'in kapası (Fleiss, 1971), pearson korelasyon katsayısı, kendall uyuşum katsayısı ve ortalamaların karşılaştırılması gibi farklı teknikler kullanılabilmektedir.

2.1.6.1.1.1. Basit Uyum Yüzdesi

Basit uyum yüzdesi, bir performansa yönelik olarak iki ya da daha fazla sayıda puanlayıcı tarafından verilen puanlar arasındaki uyumun yüzdesidir. Puanlayıcı içi güvenirliğin hesaplanması durumunda ise, basit uyum yüzdesi bir puanlayıcının aynı performansa ilişkin iki farklı zamanda yaptığı puanlamalar arasındaki uyumu göstermektedir (Goodwin, 2001). Basit uyum yüzdesini hesaplamak için iki puanlayıcının uyum içerisinde olduğu puanlamaların sayısı toplam puanlama sayısına bölünmektedir. Bu işlem sonucunda elde edilen değer puanlayıcı güvenirlik katsayısı olarak kabul edilmektedir. Puanlayıcı güvenirliğine ilişkin bu katsayı, 0 ile 1 arasında değişen değerler almaktadır (Graham, Milanowski, Miller, 2012). Basit uyum yüzdesi tekniğinin en önemli avantajı, hesaplanmasının ve

anlaşılmasının kolay olmasıdır. Bir diğer avantajı ise sınıflama, sıralama, aralık ve oran gibi tüm veri türleri için uygun olmasıdır (Goodwin, 2001). Bununla birlikte, bu teknikte puanlayıcılar arasında şans sonucu gerçekleşebilecek uyumun dikkate alınmaması, basit uyum yüzdesi tekniğinin bir dezavantajı olarak görülmektedir (Banerjee, Capozzoli, McSweeney ve Sinha, 1999). Puanlayıcıların rastgele (random) puanlama yapmaları durumunda bile puanlayıcılar arasında bir uyumun olacağı ifade edilmektedir. Puanlayıcıların rastgele puanlama yapmaları halinde puanlayıcılar arasında gözlenebilecek bu uyum şans uyumu olarak tanımlanmaktadır (Viera ve Garrett, 2005). Bu durum, puanlayıcı güvenirliği belirlenirken, puanlayıcılar arasında şans sonucu oluşabilecek uyumun göz önünde bulundurulmasını gerekli kılmaktadır. Ancak basit uyum yüzdesi tekniğinde, puanlayıcılar arasında şansa bağlı olarak oluşabilecek uyum dikkate alınmamaktadır. Cohen (1960), basit uyum yüzdesine ilişkin bu problemin aşılabilmesi için kappa istatistiğini ileri sürmüştür.

2.1.6.1.1.2. Cohen'in Kappası

Cohen'in kappası puanlayıcılar arasında gözlenen uyumun, şans ile gerçekleşmesi beklenen uyumdan ne kadar yüksek olduğunu açıklamak üzere ileri sürülmüştür (Birkimer ve Brown, 1979). Tıpkı basit uyum yüzdesi tekniği gibi kappa istatistiği de puanlayıcılar arasındaki uyumun büyüklüğüne ilişkin nicel bir ölçü vermektedir (Viera ve Garrett, 2005). Ancak, basit uyum yüzdesi tekniğinden farklı olarak Cohen'in kappası puanlayıcılar arasında gözlenen uyumun bir kısmının şanstan kaynaklandığını kabul etmektedir (Banerjee, Capozzoli, McSweeney ve Sinha, 1999). Bu nedenle, kappa istatistiği hesaplanırken şanstan kaynaklanan uyuma yönelik olarak bir düzeltme işlemi uygulanmaktadır (Vanbelle ve Albert, 2009). Kappa istatistiği aşağıda verilen eşitlik yardımıyla hesaplanmaktadır. Bu eşitlikte, P_0 gözlenen uyumu, P_c ise şans ile gerçekleşmesi beklenen uyumu göstermektedir.

$$\text{Kappa} = \frac{Po - Pc}{1 - Pc}$$

Kappa istatistiği, gözlenen uyum ile şans sonucu gerçekleşmesi beklenen uyum arasındaki farkı, "-1 ile 1" arasında kalan bir ölçek üzerinde standartlaştırmaktadır (Viera ve Garrett, 2005). Bir başka deyişle, kappa istatistiği teorik olarak -1 ile 1 arasında değişen değerler almaktadır. Negatif kappa değerleri puanlayıcılar arasında gözlenen uyumun şans ile gerçekleşmesi beklenen uyumdan

daha düşük olduğunu gösterir (Warrens, 2011). Ancak, birçok makul puanlayıcı en azından şans ile gerçekleşebilecek kadar uyumlu puanlamalar yapacağı için kappa istatistiğine ilişkin negatif değerler dikkate alınmamaktadır (David, 2008; Wynd, Schmidt ve Schaefer, 2003). Dolayısıyla, -1 ile 1 arasında değişen teorik yapıdan farklı olarak, uygulamada kappa istatistiğinin 0 ile 1 arasında değiştiği kabul edilmektedir (Goodwin, 2001). Kappa değerinin 0'a eşit olması puanlayıcılar arasında gözlenen uyumun şans ile gerçekleşmesi beklenen uyuma eşit olduğu anlamına gelmektedir. Kappa istatistiğinin 1'e eşit olması ise, puanlayıcılar arasındaki uyumun mükemmel olduğunu göstermektedir (Schuster, 2002).

Cohen (1960) tarafından ortaya atılan kappa istatistiğinde şans ile gerçekleşebilecek uyum için düzeltme işleminin uygulanması bu tekniğin en önemli avantajı olarak görülmektedir. Diğer taraftan puanlayıcılar arası tüm uyumsuzlukların eşit olmadığı durumlarda, kappa istatistiği (Cohen, 1960) uyumsuzluğun derecesini yansıtmada yetersiz kalmaktadır. Kappa istatistiğine ilişkin bu yetersizlik, araştırmacıları puanlayıcılar arası güvenirliği belirlemede kullanılacak daha kapsamlı bir istatistik arayışına yöneltmiştir (Schuster, 2004). Bu arayışların bir ürünü olarak, Cohen (1968) tarafından ağırlıklandırılmış kappa istatistiği önerilmiştir.

2.1.6.1.1.3. Ağırlıklandırılmış Kappa (Weighted Kappa)

Bazen iki puanlayıcı arasında tespit edilen uyuşmazlıkların bir kısmı diğerlerine göre daha ciddi olabilmektedir. Bu durumda, Cohen (1968) tarafından kappa istatistiğinin bir uzantısı olarak ileri sürülen ağırlıklandırılmış kappa istatistiğinin kullanılması önerilmektedir. Ağırlıklandırılmış kappa istatistiğinde, puanlayıcılar arasında tespit edilen uyuşmazlıkların derecesine bağlı olarak, bu uyuşmazlıklar hakkında kısmi ağırlıklandırma yapılmaktadır. Örneğin, iki puanlayıcının *çok yararlı, biraz yararlı, etkisiz, biraz zararlı ve çok zararlı* şeklinde beş kategorili bir puanlama aracı ile puanlama yaptığını düşünelim. Burada, puanlayıcılardan biri tarafından *çok yararlı*, diğeri tarafından *biraz yararlı* şeklinde yapılan iki puanlama arasındaki fark, *çok yararlı* ve *çok zararlı* şeklinde yapılan iki puanlama arasındaki uyumsuzluk kadar önemli değildir. Bu gibi durumlarda ağırlıklandırılmış kappa, kategoriler arasındaki uzaklıklara bağlı ağırlıklandırmalar uygulamaktadır. Böylelikle, puanlayıcılar arasındaki uyuşmazlığın derecesinin dikkate alınmasını sağlamaktadır (Viera ve Garrett, 2005).

Kappa (Cohen, 1960) ve ağırlıklandırılmış kappa (Cohen, 1968), aynı duruma ilişkin puanlama yapan iki puanlayıcı arasındaki uyumun belirlenmesi amacıyla kullanılan istatistiklerdir. Ancak bazen, aynı birey ya da nesne hakkında değerlendirme yapan puanlayıcıların sayısı ikiden fazla olabilmektedir. Fleiss (1971), puanlayıcı sayısının ikiden fazla olduğu durumlarda kullanılmak üzere, kappa istatistiğinin bir uzantısı olan "*çoklu puanlayıcılar için kappa*" istatistiğini önermiştir.

2.1.6.1.1.4. Fleiss'in Kappası

Fleiss (1971), *k* sayıda (*k*>2) gözlemcinin puanlamaları arasındaki tutarlığın belirlenmesi için kappa istatistiğinin bir uzantısı olan "çoklu puanlayıcılar için kappa istatistiğini" ortaya atmıştır. Burada, *n* bireyin her biri, birbirini dışlayan sınıflama kategorileri kullanılarak *k* sayıda puanlayıcı tarafından bağımsız olarak puanlanmaktadır (Banerjee vd., 1999). Fleiss'in (1971) kappası, tıpkı Cohen'in kappası (Cohen, 1960) ve ağırlıklandırılmış kappa (Cohen, 1968) istatistiklerinde olduğu gibi puanlayıcılar arasındaki uyumun şans eseri gerçekleşmiş olabileceğini hesaba katmaktadır. Bundan dolayı, Fleiss'in kappasının basit yüzdesi tekniğine göre daha güçlü sonuçlar ürettiği kabul edilmektedir (Özder, 2012).

2.1.6.1.1.5. Pearson Momentler Çarpımı Korelasyonu

Puanlayıcı güvenirliğinin belirlenmesinde yaygın olarak kullanılan tekniklerden biri de, iki farklı puanlayıcının yaptığı puanlamalar arasındaki korelasyon katsayısının hesaplanmasıdır. Bu amaçla genel olarak pearson momentler çarpımı korelasyonundan yararlanılmaktadır. Pearson korelasyonu, iki puanlayıcının yaptığı puanlamalar arasındaki tutarlığı gösteren bir analizdir. Bu analiz, puanlayıcıların ölçülmek istenen özellik açısından bireyler arasında yaptıkları sıralamanın tutarlığına ilişkin bir ölçü vermektedir (LeBreton ve Senter, 2008). Bir başka deyişle, pearson korelasyonu ortalamadan bağımsız olarak hesaplanmaktadır. Bu nedenle, korelasyon analizi sonucunda elde edilen değer puanların denkliğine yönelik bir bilgi vermemektedir. Yalnızca puanlar arasındaki göreli sıralamanın ne derece denk olduğunu göstermektedir (Goodwin, 2001; LeBreton ve Senter, 2008). Dolayısıyla pearson korelasyon katsayısı, basit uyum yüzdesi ya da kappa istatistiğinden farklı olarak puanlar arasındaki mutlak uyumun değil; göreli uyumun bir göstergesidir (Graham, Milanowski ve Miller, 2012).

Puanlayıcı güvenirliğinin belirlenmesinde korelasyon katsayının kullanılmasının çeşitli avantajları ve dezavantajları bulunmaktadır. En önemli avantajlarından biri, korelasyon katsayısının hesaplanmasının ve yorumlanmasının kolay oluşudur. Analiz sonuçları yorumlanırken, korelasyon katsayısının anlamlılığından ziyade büyüklüğü dikkate alınmalıdır (Goodwin, 2001). Çünkü korelasyon katsayının anlamlılığı örneklem büyüklüğünden önemli ölçüde etkilenmektedir (Baykul ve Güzeller, 2013). Örneklemi oluşturan katılımcı sayısı, null (yokluk) hipotezinin reddedilmesinde belirleyici bir rol oynamaktadır (Goodwin, 2001). Yani küçük örneklemlerde bir şey ifade etmeyen ve anlamlı çıkmayan korelasyon katsayıları büyük örneklemlerde anlamlı çıkabilmektedir (Akbulut, 2010). Örneğin; ölçmede güvenirlik hesaplamalarında .45 gibi bir değer sıfırdan manidar düzeyde büyük çıkabilir. Fakat böyle bir değer ölçme sonuçlarının güvenirliği için yeterli görülmemektedir (Baykul ve Güzeller, 2013). Korelasyon katsayının ortalamadan bağımsız olarak hesaplanan bir istatistik olması ve puanlamalar arasındaki mutlak uyumu dikkate almaması, bu tekniğin en önemli dezavantajıdır. Korelasyon katsayısı yalnızca birlikte değişimin bir ölçüsüdür. Örneğin, biri oldukça cömert diğeri oldukça katı puanlamalar yapan iki farklı puanlayıcı için hesaplanan korelasyon katsayısı oldukça yüksek çıkabilmektedir (Hux, Sanger, Reid ve Maschka, 1997). Burada korelasyon katsayısının yüksek olması iki puanlayıcının değerlendirdikleri bireyler için yaptıkları sıralamanın benzer olduğu anlamına gelmektedir. Ancak, puanlayıcıların yaptıkları puanlamalar arasındaki mutlak uyum hakkında bir bilgi vermemektedir. Bu nedenle, puanlayıcı güvenirliğinin belirlenmesinde korelasyon katsayısının kullanılması durumunda, ortalamaların da ayrıca karşılaştırılması gerekmektedir. Ortalamalar karşılaştırılarak, puanlayıcıların yaptıkları puanlamalar arasındaki farkın anlamlılığı belirlenebilir (Goodwin, 2001). Puanlayıcı güvenirliğinin belirlenmesinde korelasyon katsayının kullanılmasının bir diğer dezavantajı değerlendirilen bireylerin puanlarındaki değişimin düşük olması halinde, korelasyon katsayısının yanıltıcı olabilmesidir (Graham, Milanowski ve Miller, 2012).

2.1.6.1.1.6. Kendall Uyuşum Katsayısı

İkiden fazla puanlayıcının belirli bir performansı puanlaması durumunda puanlayıcılar arası güvenirliğin belirlenmesinde kullanılan yöntemlerden biri Kendall uyuşum (konkordans) katsayısıdır. Kendall uyuşum katsayısı (W), korelasyona

dayalı bir güvenirlik belirleme yöntemlerinden biridir ve parametrik olmayan bir istatistiksel tekniktir (Howell, 2002). Kendall uyuşum katsayısı; sıralama ölçekli veriler üzerinden hesaplanması, puanlayıcılar arası göreli uyumu yansıtması ve parametrik olmayan bir istatistiksel teknik olması yönüyle sperman sıra farkları korelasyonu ile benzerlik göstermektedir. Ancak, sperman korelasyonu iki puanlayıcının söz konusu olduğu durumlarda kullanılırken; kendall uyuşum katsayısı üç ya da daha fazla sayıda puanlayıcı olması durumunda kullanılmaktadır (Kraska-Miller, 2014). Kendall uyuşum katsayısı 0 ile 1 arasında değişen değerler almaktadır (Clark-Carter, 2010). Kendall uyuşum katsayının formülü aşağıda verilmiştir (Tavşancıl, 2010).

$$W = \frac{s}{1/12\, k^2(N^3 - N)}$$

Bu formüldeki *s*, R_j'nin ortalamasından gözlenen sapmaların karelerinin toplamını göstermektedir. Dolayısıyla, *k* puanlayıcı ve *N* puanlanan görev ya da madde sayısı olmak üzere; *s* aşağıdaki formül yardımıyla hesaplanmaktadır.

$$s = \Sigma(R_j - \frac{R_j}{N})^2$$

2.1.6.1.1.7. Ortalamaların Karşılaştırılması

Puanlayıcılar arası güvenirliğin belirlenmesinde kullanılan tekniklerden biri, ortalamalar arasındaki farkın anlamlılığının test edilmesidir. Ortalamalar arasındaki farkın anlamlılığı sınanırken, iki puanlayıcı olması durumunda ilişkili örneklem *t*-testi, ikiden fazla sayıda puanlayıcı olması durumunda ise tekrarlı ölçümler ANOVA kullanılmaktadır (Goodwin, 2001). Puanlayıcılar arası güvenirliğin belirlenmesinde korelasyon katsayının tek başına kullanılması yeterli görülmediği gibi, yalnızca ortalamalar arasındaki farkın test edilmesi de yeterli olmamaktadır. Gerek ilişkili örneklem *t*-testi, gerekse de tekrarlı ölçümler ANOVA örneklem büyüklüğüne oldukça duyarlı istatistiklerdir. Dolayısıyla, büyük örneklemler ile çalışılması halinde, pratikte anlamlı olmayan *t* veya *F* testi değerleri istatistiksel olarak anlamlı çıkabilmektedir (Akbulut, 2010). Bu durum, tespit edilen farkın önemli olduğu izlenimini uyandırarak, araştırmacıların yanılgıya düşmesine neden olabilmektedir. Ortalamalar arasındaki farkın anlamlılığının test edilmesinin yanı sıra korelasyon katsayının büyüklüğünün de dikkate alınması böyle bir yanılgıya düşme olasılığını azaltabilir.

Puanlayıcı güvenirliğinin belirlenmesinde, klasik test kuramına dayalı bu yöntemlerden hiç biri diğerlerine göre daha iyi değildir. Bu nedenle, birden fazla yöntemin bir arada kullanılması önerilmektedir (Graham, Milanowski ve Miller, 2012). Ancak farklı yöntemler bir arada kullanılsa bile, klasik test kuramına göre yapılan güvenirlik kestirimlerine ilişkin birtakım sınırlılıkların üstesinden gelmek mümkün değildir. İlk olarak klasik test kuramına göre elde edilen güvenirlik kestiriminde, ölçmeye karışan tek bir hata kaynağı dikkate alınmaktadır. Aynı anda birden fazla hata kaynağının göz önünde bulundurulmaması ve bu hata kaynakları arasındaki etkileşimin tespit edilememesi, klasik test kuramının sınırlılıkları arasında yer almaktadır. İkinci olarak, klasik test kuramı, tüm hataların tesadüfî ve tek boyutlu olduğu esasına dayanmaktadır. Bu nedenle, klasik test kuramına göre yapılan güvenirlik kestirimleri sistematik hata ile tesadüfi hatayı birbirinden ayıramamaktadır (Haiyang, 2010). Son olarak klasik test kuramında, tüm katılımcılar için tek bir standart hata belirlenmekte (Weir, 2005) ve puanlayıcı ya da performansı değerlendirilen bireyler için bireysel düzeyde bilgi vermek yerine grup düzeyinde bilgiler sunmaktadır (Barkaoui, 2008). Klasik test kuramının söz konusu sınırlılıkları, araştırmacıları bu sınırlılıkların aşılmasına imkân tanıyacak farklı model arayışlarına yöneltmiştir. Bu arayışlar neticesinde, madde tepki kuramına dayalı bir model olan Çok Yüzeyli Rasch Modeli (Linacre, 1989) ileri sürülmüştür.

2.1.6.1.3. Madde Tepki Kuramı ve Çok Yüzeyli Rasch Modeli

2.1.6.1.3.1. Madde Tepki Kuramı

Klasik test kuramına dayalı modeller ve yöntemler, eğitimde ve psikolojide kullanılan ölçme araçlarının hazırlanmasında ve bu ölçme araçlarından elde edilen puanların yorumlanmasında, uzun yıllardır test uzmanlarına hizmet etmektedir (Hambleton, Swaminathan ve Rogers, 1991). Klasik test kuramının matematiksel olarak hesaplanması kolaydır (Haiyang, 2010). Ayrıca, klasik test kuramı karşılanması kolay varsayımlara sahiptir. Bu nedenle, klasik test kuramı birçok test durumuna rahatlıkla uygulanabilmektedir (Hambleton ve Jones, 1993). Ancak, ölçme araçlarının geliştirilmesinde ve değerlendirilmesinde klasik yöntemlerin kullanılması, bir takım sınırlılıkları beraberinde getirmektedir. Ölçme aracında yer alan maddelerin güçlük ve ayırt edicilik gibi karakteristik özelliklerinin testin uygulandığı gruba ve ölçme aracının uygulandığı bireylerin yeteneklerinin ölçme aracında yer alan

maddelerin parametrelerine bağımlı olması klasik test kuramının en önemli sınırlılıklarıdır (Hambleton, 1987; Hambleton, 1995). Örneğin, bireylerin matematik başarılarını ölçmek için iki farklı matematik testinin uygulandığını varsayalım. Bu durumda, bireyler için farklı matematik testleri kullanılarak yapılan yetenek kestirimleri farklı olacaktır. Benzer şekilde, herhangi bir matematik testi iki ayrı gruba uygulandığında, testte yer alan maddeler için iki gruptan elde edilen güçlük indeksleri farklılık gösterecektir. Testteki maddelerin güçlük indeksleri matematik başarısı yüksek bir gruptan elde edilen verilere dayalı olarak hesaplandığında yüksek; matematik başarısı düşük bir örneklemden elde edilen verilere dayalı olarak belirlendiğinde ise, düşük çıkacaktır.

Klasik test kuramının bu sınırlılıklarını gidermeye yönelik olarak, 1950'li yıllarda, Frederic Lord ve George Rasch'ın çalışmaları sonucunda, Örtük Özellikler Teorisi olarak da bilenen Madde Tepki Kuramı geliştirilmiştir (Domino ve Domino, 2006). Madde tepki kuramı, bir testin ölçtüğü özellik ile bireylerin yetenek düzeyi arasındaki ilişkiyi (DeMars, 2010) matematiksel modellere dayalı bir dizi işlem ile açıklamaya çalışan bir kuramdır (Urbina, 2004). Madde tepki kuramı ile bir maddenin karakteristik özellikleri maddeyi yanıtlayan bireylerden bağımsız olarak elde edilebildiği gibi, bireyin yetenek düzeyi de yanıtladığı maddelerden bağımsız olarak tahmin edilebilmektedir (Borlotti, Tezza, Andrade, Bornia ve Junior, 2013; Ostini ve Nering, 2006). Ayrıca, madde tepki kuramı, tüm bireyler için tek bir standart hata veren klasik test kuramının aksine, ölçme aracının uygulandığı her birey için ayrı bir standart hata üretmektedir (Embretson ve Reise, 2000).

Madde tepki kuramı; üç parametreli, iki parametreli ve bir parametreli model olmak üzere üç farklı modelden oluşmaktadır. Bu yönüyle madde tepki kuramı tek bir modele dayandırılan bir kuramdan çok; bir modeller ailesini yansıtmaktadır (DeVellis, 2003). Üç parametreli model; madde güçlüğü, madde ayırt ediciliği ve şans parametrelerinden meydana gelmektedir. Madde güçlük parametresi (b) bir maddeyi doğru yanıtlama olasılığının .50 olduğu yetenek düzeyini (θ) ifade etmektedir. Madde güçlük indeksinin ranjı teorik olarak $-\infty$ ile $+\infty$ arasında değişmektedir. Bununla birlikte, madde güçlük indeksinin pratikteki değerleri ± 3 aralığında yer almaktadır (Baker, 2001). Negatif b değerleri maddenin kolay, pozitif b değerleri maddenin zor ve sıfıra yakın b değerleri maddenin orta güçlükte olduğunun işareti kabul edilir (Atılgan, 2009). Madde ayırt edicilik parametresi (a), maddenin, ölçülen özellik açısından farklı yetenek düzeylerindeki bireyleri

birbirinden ne kadar iyi ayırt edebildiğinin bir göstergesidir. Madde ayırt edicilik parametresi, madde karakteristik eğrisinin kırılma noktasındaki (θ=b) eğimi olarak da tanımlanmaktadır (Baker, 2001). Bu eğim hesaplanarak, yetenek düzeyi artıkça maddeyi doğru yanıtlama olasılığının nasıl değiştiği belirlenebilmektedir (DeMars, 2010). Şans parametresi (c) ise, düşük yetenek düzeyindeki bir katılımcının güçlük açısından orta ya da yüksek düzeyde olan bir maddeyi tahminle doğru yanıtlama olasılığını göstermektedir (Haiyang, 2010). Bu tanımdan anlaşılacağı üzere, şans parametresi yetenek düzeyinin bir fonksiyonu olarak değişmemektedir. Bu nedenle, düşük ve yüksek yetenek düzeyindeki katılımcıların bir maddeyi şans ile doğru cevaplandırma olasılıkları aynıdır. Şans parametresi teorik olarak 0 ile 1 arasında değişen değerler alabilmektedir. Ancak, şans parametresine ilişkin .35'in üzerindeki değerlerin kabul edilebilir olmadığı düşünülmektedir. Dolayısıyla, şans parametresi için uygulamada $0 \leq c \leq .35$ aralığı dikkate alınmaktadır (Baker, 2001).

Madde tepki kuramın bir diğer modeli, madde güçlük ve madde ayırıcılık parametrelerinden oluşan iki parametreli modelidir. İki parametreli model, üç parametreli modelden farklı olarak şans parametresini içermemektedir (Reckase, 2009). Bundan dolayı, iki parametreli model, üç parametreli modeldeki şans parametresinin sıfıra eşit olduğu bir model olarak düşünülebilir (Fox, 2010). Üç parametreli modelde madde karakteristik eğrisinin en düşük noktası şans parametresinin değerine eşittir. İki parametreli modelde ise, madde karakteristik eğrisinin en düşük değeri sıfırdır (Baker, 2001).

Madde tepki kuramının bir diğer modeli, bir parametreli modeldir. Bir parametreli model, yalnızca madde güçlük parametresini içermektedir. Bir parametreli modelde tüm maddelerin madde ayırıcılık indekslerinin birbirine eşit olduğu kabul edilmektedir. Dolayısıyla, bir parametreli modelde madde karakteristik eğrisinin şekli tüm maddeler için aynı olmaktadır. Bu modelde, madde karakteristik eğrileri arasındaki tek fark, yatay eksen (θ) boyunca eğrinin sağ ve sol tarafında kalan bölgedir (Harvey ve Hammer, 1999). Diğer bir ifadeyle, bir parametreli modelde, farklı maddelerin madde karakteristik eğrileri, yalnızca dikey eksene olan uzaklıkları yönüyle birbirinden ayrılmaktadır.

Bir parametreli model, bu modeli geliştiren Danimarkalı matematikçi ve istatistikçi George Rasch'ın (1960, 1980) adıyla anılmakta ve Rasch modeli olarak da ifade edilmektedir. Rasch modeli, ilk olarak standart başarı testleri ya da çoktan seçmeli testler gibi doğru/yanlış şeklinde puanlanabilen iki kategorili (dichotomous)

ölçme araçları için geliştirilmiştir (Sebok, Luu ve Klinger, 2013). Rasch modelinin ortaya atılmasından sonra birçok araştırmacı Rasch modeli ile ilgili çalışmalar yapmıştır. Bu çalışmalar sonucunda, Rasch modelinin farklı uzantıları ileri sürülmüştür. Andrich (1978) temel Rasch modelini Likert tipi ölçek verilerinin analizinde kullanılabilecek şekilde genişletmiştir. Andrich'in önerdiği ve temel Rasch modelinin genişletilmiş hali olan bu model *sıralama ölçekli model* (rating scale model) olarak isimlendirilmiştir (Engelhard, 2005). Andrich'in (1978) çalışmalarının bir uzantısı olarak, Masters (1982) *kısmi puanlama modelini* (partial credit model) geliştirmiştir. Kısmi puanlama modeli, açık uçlu soruları analiz etmek amacıyla geliştirilmiş bir modeldir (Koskey, 2009). Bu modelde, maddeler için doğru/yanlış şeklinde bir değerlendirme yapmak yerine; çözüm sürecinde takip edilen adımlar göz önünde bulundurularak kısmi puanlamalar yapılabilmektedir (Cagnone ve Ricci, 2005). Bununla birlikte, gerek Andrich'in (1978) önerdiği sıralama ölçekli model gerekse de Masters'ın (1982) kısmi puanlama modeli, yetenek ve madde güçlüğü olmak üzere iki yüzey içermektedir. Linacre (1989), Masters'in (1982) kısmi puanlama modelini bir adım ileriye taşıyarak, puanlayıcı etkisi gibi farklı değişkenlik kaynaklarını da modele dâhil etmiştir (Farrokhi ve Esfandiari, 2011). Ölçme sonuçlarını etkileyebilecek farklı değişkenlik kaynaklarının kısmi puanlama modeline dâhil edilmesiyle, Çok Yüzeyli Rasch Modeli olarak ifade edilen yeni bir modele ulaşılmıştır (Mulqueen, Baker ve Dismukes, 2000).

2.1.6.1.3.2. Çok Yüzeyli Rasch Modeli

Temel Rasch modelinde, test puanlarını etkileyebilecek değişkenlik kaynakları olarak, yalnızca bireylerin yetenek düzeyleri ile maddelerin güçlük düzeyleri esas alınmaktadır. Çok yüzeyli Rasch modelinde ise, test puanlarını etkileyebileceğine inanılan tüm değişkenlik kaynakları göz önünde bulundurulmaktadır (Baird, Hayes, Johnson, Johnson ve Lamprianou, 2013). Bu kapsamda, çok yüzeyli Rasch modeli, potansiyel olarak test puanlarını etkileme olasılığına sahip tüm değişkenlik kaynaklarının dikkate alınmasını sağlayan bir model olarak tanımlanabilir (Kim, Park ve Kang, 2012). Çok yüzeyli Rasch modelinde, test puanlarını etkileme olasılığı bulunan tüm değişkenlik kaynaklarının analize dâhil edilmesinin yanı sıra, bu değişkenlik kaynakları arasındaki etkileşimler de belirlenebilmektedir (Abu Kassim, 2007). Örneğin; test puanlarında değişkenliğe neden olabilecek faktörler arasında yer alan performans görevi ve yetenek düzeyi

değişkenleri arasında anlamlı bir etkileşim olup olmadığı belirlenebilir. Bu sayede, bir grup bireyin belli bir görevde, sistematik olarak diğer görevlerden farklı bir performans gösterip göstermediği ortaya konulabilir. Benzer şekilde, puanlayıcıların bireylerin performanslarını puanlarken, bir ya da bir grup katılımcıya diğerlerinden daha yüksek puan verme eğiliminde olup olmadığı çok yüzeyli Rasch modeli ile belirlenebilir. Daha genel bir ifadeyle, çok yüzeyli Rasch modeli ile "*birey*×*madde*", "*puanlayıcı*×*birey*" ve "*puanlayıcı*×*madde*" gibi test puanlarını etkileme olasılığı bulunan değişkenlik kaynaklarının ikili kombinasyonları arasındaki etkileşimler belirlenebilmektedir. Çok yüzeyli Rasch modeli kullanılarak yüzeyler arasındaki etkileşimin belirlenebilmesi, çok yüzeyli Rasch modelini yanlılık analizlerinin tespitinde de kullanılabilecek bir model haline getirmektedir (Sudweeks, Reeveb ve Bradshawc, 2004).

Çok yüzeyli Rasch modelinde, test puanlarını etkileme olasılığına sahip değişkenlik kaynaklarının her biri yüzey olarak adlandırılmaktadır (Koons, 2008; Sudweeks, Reeveb ve Bradshawc, 2004). Örneğin; bir matematik performansının ölçülmesi için, öğrencilere dört farklı görev verilmiş ve bu görevlerin her biri iki farklı puanlayıcı tarafından bağımsız olarak puanlamış olsun. Burada; öğrenciler, görevler ve puanlayıcılar ölçme sonuçlarını etkileyebilecek değişkenlik kaynaklarıdır. Dolayısıyla; öğrenci yüzeyi, madde yüzeyi ve puanlayıcı yüzeyi şeklinde üç yüzeyli bir model söz konusu olmaktadır. Çok yüzeyli Rasch modelinde, puanlayıcı ile ilgili faktörlerin bireyin puanlarında değişkenliğe neden olabilecek bir yüzey olarak işlem görmesi, bu modeli öznel olarak puanlanan performans görevleri için uygun hale getirmektedir (Mulqueen, Baker ve Dismukes, 2000). Çok yüzeyli Rasch modeli kullanılarak, puanlayıcılar arasında tutarsızlık olup olmadığı, puanlayıcıların katılık ve cömertlikleri yönüyle farklılık gösterip göstermediği tespit edilebilmektedir (Matsuno, 2009). Yine çok yüzeyli Rasch modelinden elde uyum istatistikleri incelenerek, ranj sınırlaması ve halo etkisi gibi puanlayıcı etkileri belirlenebilmektedir. Çok yüzeyli Rasch modeli puanlayıcı etkilerinin belirlenmesi ile sınırlı kalmamaktadır. Çok yüzeyli Rasch modelinde, puanlayıcı farklılıklarına yönelik düzeltme işlemleri de uygulanmaktadır (Abu Kassim, 2007). Bu sayede puanlayıcı farklılıkları kontrol altına alınabilmektedir (Linacre, 1989; Linacre, Engelhard, Tatum ve Myford, 1994).

Çok yüzeyli Rasch modelinin en önemli avantajlarından biri; birey, madde ya da puanlayıcı gibi değişkenlik kaynakları hakkında grup düzeyinde bilgiler veren klasik test kuramının (Barkaoui, 2008) aksine, değişkenlik kaynaklarının her bir bileşeni için araştırma yapmaya imkân tanımasıdır (Yue, 2011). Çok yüzeyli Rasch modelinin, modelde yer alan her bir yüzeyin her bir elemanına ilişkin bilgi vermesi (Kozaki, 2004), bir yüzeyi oluşturan bileşenlerden herhangi birinin uyumsuz olup olmadığını belirlemeyi olanaklı hale getirmektedir. Örneğin, çok yüzeyli Rasch modeli kullanılarak, diğer puanlayıcılara göre daha katı ya da daha cömert puanlamalar yapan bir puanlayıcı olup olmadığı belirlenebilmektedir. Benzer şekilde, çok yüzeyli Rasch modeli sayesinde, yanıtları arasında tutarsızlık olan bireyler bulunup bulunmadığı ya da testti oluşturan maddelerden herhangi birinin testteki diğer maddelere kıyasla daha zor olup olmadığı ortaya konulabilmektedir (Kozaki, 2004; Mulqueen, Baker ve Dismukes, 2000).

Rasch modelinin bir diğer avantajı; madde güçlüğü, bireyin yetenek düzeyi ve puanlayıcılar gibi analize dâhil edilen bütün yüzeyleri ayrı ayrı kalibre etmesine rağmen, bunları ortak bir ölçek üzerine yerleştirmesidir (O'Neill ve Lunz, 1996). Buna göre, her bir yüzeyin parametreleri diğer yüzeylerden bağımsız olarak tahmin edilmektedir (Kramer, Bowyer, Kielhofner, O'Brien ve Barbosa, 2009). Örneğin; puanlayıcı etkisine ilişkin yapılan tahminler, örneklemin ya da testin karakteristik özelliklerinden etkilenmemektedir. Ancak, analiz sonucunda, birey, madde ya da puanlayıcı gibi modelde yer alan bütün yüzeyler ortak bir ölçek üzerine yerleştirilmektedir (Kim, Park ve Kang, 2012). Bu ölçek eşit aralık düzeyinde olup ölçekte kullanılan olasılık birimi logittir. Logit ölçeği $-\infty$ ile $+\infty$ aralığında değişen değerler almaktadır. Ölçekte yer alan değerler, modeldeki farklı yüzeyler için değişik şekillerde yorumlanmaktadır. Puanlayıcı yüzeyi için pozitif değerler puanlayıcıların katı puanlamalar yaptığına işaret ederken; negatif değerler puanlayıcıların puanlama işleminde cömert davrandığını göstermektedir (Randall ve Engelhard, 2009). Birey yüzeyi söz konusu olduğunda, pozitif değerler bireyin yetenek düzeyinin yüksek olduğu anlamına gelirken; negatif değerler bireyin yetenek düzeyinin düşük olduğunu yansıtmaktadır. Madde yüzeyi için ise, pozitif logit değerleri maddenin güçlük düzeyinin yüksek olduğunu gösterirken; negatif logit değerleri maddenin güçlük düzeyinin düşük olduğu anlamına gelmektedir. Çok yüzeyli Rasch modelinde bütün yüzeylerin ortak bir ölçek üzerine yerleştirilmesi, modeldeki tüm yüzeylerin birbirleri ile karşılaştırılabilmesini sağlamaktadır (Schaefer, 2008).

2.1.6.1.3.2.1. Çok Yüzeyli Rasch Modelinin Varsayımları

Çok yüzeyli Rasch modeli, temel Rasch modelinin bir uzantısıdır. Temel Rasch modeli ise, madde tepki kuramındaki bir parametreli modele karşılık gelmektedir. Bu nedenle; çok yüzeyli Rasch analizleri için karşılanması gereken varsayımlar, madde tepki kuramının varsayımları ile aynıdır (Farrokhi ve Esfandiari, 2011). Bunlar; tek boyutluluk, yerel bağımsızlık ve model ile veri uyumu varsayımlarıdır (DeMars, 2010).

Tek Boyutluluk: Tüm madde tepki kuramı modellerinde olduğu gibi, temel Rasch modelinin bir uzantısı olan çok yüzeyli Rasch modeli için karşılanması gereken varsayımlardan biri ölçülen özelliğin tek boyutlu bir yapıya sahip olmasıdır. Bu varsayım, testte yer alan maddelerin tamamının tek bir özelliği ölçmeye yönelik olması anlamına gelmektedir (Harvey ve Hammer, 1999). Tek boyutluluk varsayımının ihlal edilmesi, madde parametrelerine ya da standart hatalara ilişkin yapılan tahminlerde yanlışlığa neden olmaktadır (DeMars, 2010). Dolayısıyla, tek boyutluluk, birçok araştırmacı tarafından çok yüzeyli Rasch modelin uygulanabilirliğini kısıtlayıcı bir varsayım olarak görülmektedir. Bu durum, kompozisyon yazma, kendini yabancı bir dilde ifade etme gibi farklı bileşenlerden oluşan birçok yeteneğin ölçülmesinde, çok yüzeyli Rasch modelinin uygun bir tercih olmayacağı endişesine yol açmaktadır. McNamara (1997) psikometrik ve psikolojik tek boyutluluk kavramlarını birbirinden ayırarak bu duruma açıklık getirmektedir. McNamara'ya (1996) göre, bir yapı psikolojik açıdan çok boyutlu olmasına rağmen; psikometrik açıdan tek boyutlu olabilmektedir. Örneğin; kompozisyon yazma becerisi psikolojik açıdan gramer, planlama, akıcılık, noktalama işaretlerinin kullanımı gibi çok sayıda farklı boyuttan oluşmaktadır. Bu nedenle, kompozisyon yazma becerisinin psikolojik olarak çok boyutlu olduğu söylenebilir. Ancak, sıralanan bileşenlere verilen puanlar toplanarak kompozisyon yazma becerisine ilişkin genel bir puan elde edilebilmektedir. Dolayısıyla, psikolojik açıdan çok boyutlu bir özellik olan kompozisyon yazma becerisinin psikometrik açıdan tek boyutlu olduğu söylenebilir (Barkaoui, 2008).

McNamara'nın (1996), psikolojik ve psikometrik çok/tek boyutluluk ayrımına benzer bir açıklama Reckase, Ackerman ve Carlson (1988) tarafından yapılmıştır. Reckase, Ackerman ve Carlson'a (1988) göre, bir testi oluşturan maddelerin ölçtüğü özellik, psikologlar ya da eğitimciler tarafından çok boyutlu bir yapı olarak kavramsallaştırılmasına rağmen; bu test matematiksel olarak tek boyutlu

olabilir. Örneğin; bir matematik testinde yer alan maddeler, öğrencilerin hem matematik bilgisini hem de matematik kaygısını ölçmeye yönelik olabilir. Eğer testteki maddelerin tamamı teste yer alan iki yapıyı aynı oranda ölçüyorsa, test matematiksel açıdan maddelerin ölçtüğü iki yapının birleşiminden meydana gelen tek bir yeteneği (θ) ölçmeye yönelik olacaktır (Reckase, Ackerman ve Carlson, 1988). Ayrıca, teorik olarak iki farklı yapıyı ölçen maddelerden oluşan bir testin uygulandığı grup, maddelerin ölçtüğü iki yapıdan biri açısından farklılaşırken, diğer yapı açısından farklılık göstermiyorsa test matematiksel olarak tek boyutlu çıkacaktır. Mesela, testi oluşturan maddeler matematik bilgisi ve matematik dersine yönelik motivasyon şeklinde iki farklı yapıyı ölçüyor olsun. Testin uygulandığı grup matematik dersine yönelik motivasyon açısından farklılaşmazken, matematik bilgisi açısından farklılık gösteriyorsa, teste verilen yanıtlar matematiksel olarak tek boyutlu olacaktır. Ancak, testi oluşturan maddelerin her birinde ölçülen iki yapının farklı oranlarda kombinasyonu söz konusuysa ve testin uygulandığı grup maddelerin ölçtüğü iki yapı açısından da farklılık gösteriyorsa test matematiksel olarak çok boyutlu çıkacaktır. Yukarıdaki örnek için, bazı maddeler yüksek düzeyde matematik motivasyonu fakat düşük düzeyde matematik bilgisi gerektirirken; bazı maddeler düşük düzey matematik motivasyonu fakat yüksek düzeyde matematik bilgisi gerektiriyorsa ve testin uygulandığı grup hem matematik motivasyonu hem de matematik bilgisi açısından farklılık gösteriyorsa test çok boyutlu çıkacaktır. Özetle, tek boyutluluk varsayımı incelenirken, boyutluluk kavramının bağlama ve örnekleme göre değişkenlik gösterebileceği dikkate alınmalı ve psikolojik çok boyutluluk ile psikometrik çok boyutluluk kavramları arasındaki ayrım göz önünde bulundurulmalıdır (DeMars, 2010). Tek boyutluluk varsayımının karşılanıp karşılanmadığı genellikle faktör analizi ile incelenmektedir.

Yerel Bağımsızlık: Çok yüzeyli Rasch modelinin bir diğer varsayımı yerel bağımsızlıktır. Yerel bağımsızlık belirli bir yetenek düzeyindeki öğrencinin, bir maddedeki performansının diğer maddelere bağımlı olmamasıdır (Farrokhi, Esfandiari ve Vaez Dalili, 2011). Diğer bir ifadeyle, yerel bağımsızlık yetenek düzeyi kontrol altına alındığında, bireyin farklı maddelere verdiği yanıtlar arasında ilişki olmaması anlamına gelmektedir (Reeve ve Fayers, 2005). Burada, maddeler arasındaki bağımsızlık incelenen örtük özelliğin tek bir noktasına yerleşen bireyler için tanımlanmaktadır. Maddeler arasındaki bağımsızlık, aynı yetenek düzeyinde bulunan alt gruplar için tanımlandığından, bu varsayım için yerel bağımsızlık ifadesi

kullanılmaktadır (Crocker ve Algina, 1986). Yerel bağımsızlık varsayımına göre, bireyin bir maddeye verdiği cevap; yalnızca bireyin yetenek düzeyinin bir sonucudur. Bireyin belirli bir maddeye verdiği cevap, ölçme aracında yer alan diğer maddelere verdiği cevaplardan etkilenmemektedir (Araujo, Andrade ve Bortolotti, 2009).

Bir veri seti için testte yer alan maddelerin hem aynı özelliği ölçmeye yönelik olması hem de bir maddeye verilen yanıtın diğer maddelerden bağımsız olması mantığa aykırı görünebilir. Testte yer alan tüm sorular aynı özelliği ölçmesine rağmen, bu sorular arasında bir ilişki olmaması ve öğrencilerin farklı maddelere verdikleri yanıtlar arasında anlamlı bir korelasyonun beklenmemesi bir çelişki olarak değerlendirilebilir. Aynı özelliği ölçmeye yönelik maddeler arasında anlamlı bir korelasyon olacağı ve maddelerin bir takım ortak özelliklerinin bulunacağı düşünülebilir. Yerel bağımsızlık varsayımı, bu ortak özellikler kontrol altına alındıktan sonra, maddeler arasında bir ilişki bulunmaması esasına dayanmaktadır. Tüm maddeler ölçülmek istenen örtük özellik ile ilişkili olacağından, yerel bağımsızlık varsayımı yetenek değişkeni kontrol altına alındıktan sonra test edilmelidir. Yani yerel bağımsızlık varsayımının test edilebilmesi için yetenek değişkenin kontrol altında tutulması şarttır. Bu nedenle, bu varsayım için şartlı bağımsızlık ifadesi de kullanılmaktadır. Yerel bağımsızlık, tek boyutluluk ile paralel çalışan bir varsayım olduğundan, yerel bağımsızlığın test edilmesinde de faktör analizi tekniklerinden yararlanılmaktadır (Hambleton, Swaminathan ve Rogers, 1991).

Model ile Veri Uyumu: Madde tepki kuramının temel varsayımlarından biri model ile verinin uyumlu olmasıdır (Chernyshenko, Stark ve Chan, 2001). Model ile veri arasındaki uyum tahmin edilen parametre sayısı ile ilgili bir varsayımdır (Barkaoui, 2008). Genel olarak, madde tepki kuramına dayalı analizlerde, yalnızca madde güçlük parametresini içeren bir parametreli model, madde güçlüğü ile madde ayırt ediciliği parametrelerinden oluşan iki parametreli model ya da madde güçlüğü, madde ayırt ediciliği ve şans parametrelerini içeren üç parametreli modelden hangisi veriler ile daha iyi uyum gösteriyorsa, analizlerin bu modele göre yapılması gerekmektedir (Baker, 2001). Özellikle, iki kategorili olarak puanlanan maddelerde model ile veri uyumunun değerlendirilmesi oldukça önemlidir. Bununla birlikte, performans değerlendirmede, temel Rasch modelinin bir uzantısı olan çok yüzeyli Rasch modelinin yerine kullanılabilecek alternatif bir model bulunmamaktadır (McNamara, 1996). Dolayısıyla, model ile veri arasındaki uyum yüksek olmasa da;

performans değerlendirmede bir parametreli modelin bir uzantısı olan çok yüzeyli Rasch modeli kullanılmaktadır. Çok yüzeyli Rasch modelinde, model ile veri arasındaki uyum hakkında verilecek kararlarda standartlaştırılmış artık değerleri referans alınmaktadır.

2.1.6.1.3.2.2. Çok Yüzeyli Rasch Modeli İstatistikleri

Çok yüzeyli Rasch modeli istatistikleri FACET (Linare, 2014) paket programı kullanılarak gerçekleştirilmektedir. Bu program her bir yüzey için, parametre tahminlerini, bu parametrelerin güvenirliğine ilişkin bilgi veren standart hata değerlerini ve ölçümlerin geçerliği hakkında bilgi sunan uygunluk istatistiklerini üretmektedir (Barkaoui, 2008). FACET programında ilk olarak, iterasyon işlemi uygulanmaktadır. İterasyon, verilerin program tarafından kaç defa okunduğunu göstermektedir. Uygulanacak iterasyon sayısı, veriden iyi tahminde bulunmanın ne kadar güç olduğuna bağlı olarak değişmektedir. Veriler Rasch modeli ile düşük uyum gösteriyorsa, veri matrisi ayrık alt gözlemlerden oluşuyorsa (erkek öğrencilerin A puanlayıcısı ve kız öğrencilerin B puanlayıcısı tarafından puanlanması gibi) ya da yüzeylerdeki bileşenlerin parametrelerine ilişkin dağılım çarpık veya çok modlu ise iyi tahminde bulunmak daha güç olmaktadır. İyi tahminde bulunmanın güç olduğu durumlarda gerekli olan iterasyon sayısı daha fazla olmaktadır (Linacre, 2014).

İterasyon raporundan sonra değişken haritası (variable map) sunulmaktadır. Değişken haritası, analizdeki tüm yüzeylerin bir arada değerlendirilebildiği bütüncül bir yaklaşım sağlamaktadır. Diğer bir ifadeyle, değişken haritası sayesinde, çok yüzeyli Rasch modeline dâhil edilen bütün yüzeyler görsel olarak birbirleriyle karşılaştırılabilmektedir (Kaliski vd., 2012). Değişken haritasındaki sütun sayısı, analize dâhil edilen yüzey sayısına bağlıdır. Bununla birlikte, analize kaç yüzey dâhil edilmiş olursa olsun değişken haritasının ilk sütununda logit ölçeği yer almaktadır. Logit ölçeğinin bulunduğu sütunun alt ucunda negatif; üst ucunda ise pozitif değerler bulunmaktadır. Değişken haritasının son sütununda ise, puanlama ölçeği yer almaktadır. Bu sütunun alt acunda puanlama ölçeğinin alt kategorileri; üst ucunda ise puanlama ölçeğinin üst kategorileri bulunmaktadır. Değişken haritasının diğer sütunlarında birey, madde ve puanlayıcı gibi analize dâhil edilen yüzeyler yer almaktadır. Bu sütunlarda; aşağıdan yukarı doğru gidildikçe, madde güçlüğü, bireylerin yetenek düzeyi ve puanlayıcıların katılığı artmaktadır (Wiseman, 2008).

FACET programından elde edilen analiz çıktılarında yer alan bir diğer tablo, kategori istatistikleridir. Bu tabloda; puanlama ölçeğinin kategorileri, hangi puanlama kategorisinin kaç defa kullanıldığını gösteren frekans değerleri, bu frekans değerlerine ilişkin yüzdeler ve yığılmalı yüzdeler yer almaktadır. Puanlama ölçeğinin etkili bir biçimde çalışabilmesi için ölçeğin her bir kategorisinde en az 10 gözlem bulunmalıdır. Kategori istatistikleri tablosunda, puanlama ölçeğinin kullanım örüntüsünü gösteren sütunlardan sonra, kalite kontrol sütunu yer almaktadır. Kalite kontrol sütununda; ortalama ölçüm, beklenen ölçüm ve uygunluk dışı istatistik değerleri bulunmaktadır. Ortalama ölçümlerin monoton olarak artması ve uygunluk dışı istatistiklerinin 2'den düşük olması, puanlama ölçeğinin etkin bir biçimde çalışması için karşılanması gereken şartlar arasındadır (Linacre, 2014).

FACET programında rapor edilen bir diğer istatistik puanlayıcı ölçüm raporlarıdır. Puanlayıcı ölçüm raporlarının sunulduğu tabloda, puanlayıcılardan her birinin yaptığı puanlamaların toplamı (total score), her bir puanlayıcı tarafından verilen puanların sayısı (total counts) ve gözlenen ortalamalar (observed average) bulunmaktadır. Gözlenen ortalamalar her bir puanlayıcının yaptığı puanlamaların ortalamasını göstermektedir. Bu ortalamalar, puanlayıcının yaptığı puanlamaların toplamının, yaptığı puanlama sayısına bölünmesiyle (total score/total counts) elde edilmektedir. Bu tabloda ayrıca, düzeltilmiş ortalamalar (fair mean average) da yer almaktadır. Düzetilmiş ortalamalar hesaplanırken, puanlayıcının puanladığı bireylerin yetenek düzeyleri ve diğer puanlayıcıların katılıkları/cömertlikleri dikkate alınarak, düzeltme işlemi uygulanmaktadır. Buna göre, gözlenen ortalamaların öğrencilerin yetenek düzeyleri ve puanlayıcılar arasındaki sapmalara göre düzeltilmesiyle düzeltilmiş ortalamalar elde edilmektedir. Düzeltilmiş ortalamaların sağ tarafındaki sütunda, her bir puanlayıcıya ilişkin logit değerleri (measure) yer almaktadır. Bu logit değerleri kullanılarak puanlayıcılar, puanlama işlemindeki katılık ve cömertlikleri açısından birbirleri ile karşılaştırılabilmektedir. Bir puanlayıcının pozitif logit değerine sahip olması katı puanlamalar yaptığını, negatif logit değerine sahip olması ise, cömert puanlamalar yaptığını göstermektedir. Tabloda, logit değerlerinin yer aldığı sütunun sağ tarafında standart hata değerleri bulunmaktadır. Standart hata değerleri, yapılan kestirimlerin doğruluğunu göstermektedir. Sonraki iki sütunda ise, uygunluk içi (Infit Mean Squares) ve uygunluk dışı (Outfit Mean Squares) istatistikleri bulunmaktadır. Bu uygunluk

istatistiklerinin .5 ile 1.5 aralığında yer alması (Linacre, 2014), uyumsuz puanlama yapan puanlayıcının bulunmadığı anlamına gelmektedir.

Çok yüzeyli Rasch analizlerinde rapor edilen diğer bir istatistik uygunluk istatistikleridir. Uygunluk istatistikleri; her bir bireyin, maddenin ya da puanlayıcının gözlenen ve beklenen değerinin birbiri ile ne kadar iyi eşleştiğini göstermektedir (Sudweeks, Reeveb ve Bradshawc, 2004). Gözlenen değer ile beklenen değer arasındaki fark artıklar (residuals) olarak tanımlanmaktadır. Her bir artık, verilerin kalitesi hakkında bir miktar bilgi vermektedir. Artık değerlerin büyük olması model ile veri arasındaki uyumun düşük olduğuna işaret etmektedir. Linacre'ye (2014) göre, standartlaştırılmış artıkların yaklaşık %5'inden fazlasının 2 değerini ve %1'inden fazlasının 3 değerini aşmaması gerekir. Bu şartın sağlanması, model ile veri arasında kabul edilebilir uyumun bulunduğunu yansıtmaktadır. Çok yüzeyli Rasch modelinde, standart artıkların karesi kullanılarak, uygunluk içi ve uygunluk dışı şeklinde ifade edilen iki uyum istatistiğine ulaşılmaktadır. Bu uygunluk istatistikleri standart artıklara göre, örneklem büyüklüğüne daha az duyarlıdır. Bundan dolayı, gözlenen değer ile beklenen değer arasındaki eşleşmenin bir ölçüsü olarak, standart artıklar yerine bu uygunluk istatistiklerinden yararlanılması önerilmektedir (Brentari ve Golia, 2008). Uygunluk içi ve uygunluk dışı istatistikleri arasındaki tek fark, uygunluk içi istatistiği hesaplanırken, maddeyi yanıtlama olasılığı %50'ye yakın olan bireylere daha yüksek bir ağırlıklandırma uygulanmasıdır (Liu, 2010). Daha açık bir anlatımla; uygunluk içi istatistiğinde, yetenek düzeyi maddenin güçlük düzeyine yakın bireyler için daha fazla; yetenek düzeyi maddenin güçlük düzeyinden uzak olan bireyler için ise daha az bir ağırlıklandırma uygulanmaktadır. Ancak, uygunluk dışı istatistiğinde, herhangi bir ağırlıklandırmaya yer verilmemektedir. Bu durum, uygunluk dışı istatistiğinin uç değerlere karşı daha hassas olmasına neden olmaktadır. Dolayısıyla, gözlenen değer ile beklenen değer arasındaki uyumun değerlendirilmesinde, daha çok uygunluk içi istatistiğinin dikkate alınması önerilmektedir (Bond ve Fox, 2001).

Uygunluk içi ve uygunluk dışı istatistikleri, 0 standart hata ve 1 beklenen değerine sahiptir. Uygunluk istatistiklerinin 1'e eşit olması, model ile veri arasında mükemmel uyum olduğunu göstermektedir. Ancak, gerçek ölçme durumları için model ile veri arasında mükemmel uyum olması genellikle imkânsızdır (Brentari ve Golia, 2008). Bu durum, uygunluk istatistiklerine ilişkin kabul edilebilir aralığın ne olduğu sorusunu akıllara getirmektedir. Bu konuda araştırmacılar arasında tam bir

uzlaşı bulunmamaktadır. Wright ve Linacre (1994), .6 ile 1.4 arasında kalan uygunluk istatistiklerini kabul edilebilir olarak nitelendirmiştir. Bu ölçüte göre, 1.5 ve üzerindeki değerler, verilerin ölçüm için uygun olmadığı şeklinde yorumlanmaktadır. Myford ve Wolfe (2003) ise, 2'ye kadar olan uygunluk istatistiklerini kabul edilebilir olarak ifade etmiştir. Myford ve Wolfe'a (2003) göre, 1.5 ile 2 arasındaki değerler verilerin ölçüm için yararlı olmadığını göstermekte; ancak, zararlı da olmadığı anlamına gelmektedir. 2'nin üstündeki uygunluk istatistikleri ise, verilerin ölçüm için zararlı olduğunu yansıtmaktadır (Sudweeks, Reeveb ve Bradshawc, 2004). Analizde yer alan herhangi bir yüzeyin herhangi bir bileşeni için hesaplanan uygunluk istatistiklerinin .6'dan düşük olması, bu bileşenin bulunduğu yüzeydeki diğer bileşenlerden daha farklı bir bilgi sağlamadığını yansıtmaktadır. Örneğin, madde yüzeyinde yer alan bazı maddeler arasında yüksek korelasyon olduğunda, uygunluk içi ve uygunluk dışı istatistikleri genellikle .6'dan düşük çıkmaktadır. Madde yüzeyi için hesaplanan bu şekildeki bir sonuç, yerel bağımsızlık varsayımının ihlal edildiğini göstermektedir (Chan, Chien, Su ve Lin, 2009). Uygunluk istatistiklerinin 2'den yüksek olması, uygunluk istatistiği değeri yüksek olan bileşenlerin, bulunduğu yüzeydeki diğer bileşenler ile aynı yapıyı ölçmediği şeklinde yorumlanmaktadır. Örneğin; herhangi bir madde için hesaplanan uygunluk istatistiğinin 2'den yüksek olması, bu maddenin ölçme aracında yer alan diğer maddeler ile aynı özelliği ölçmediğini göstermektedir. Benzer şekilde, puanlayıcı yüzeyi için hesaplanan uygunluk istatistiklerinin kabul edilebilir aralığın dışında kalması puanlama işlemine, puanlayıcı hatalarının karıştığını göstermektedir (Engelhard, 2011). Çok yüksek ya da çok düşük uygunluk istatistikleri puanlayıcıların tutarsız ya da aşırı tutarlı puanlamalar yaptığını göstermektedir (Moore, 2009). Mesela, herhangi bir puanlayıcı için hesaplanan uygunluk içi ve uygunluk dışı istatistiklerinin düşük olması (.6'dan düşük olması) puanlayıcıların maddeleri birbirinden bağımsız olarak puanlayamadığı anlamına gelmekte ve ölçme işlemine halo etkisinin karıştığını göstermektedir (Engelhard, 2011).

FACET programı, analize dâhil edilen her bir yüzey için ayrı bir güvenirlik katsayısı üretmektedir. Çok yüzeyli Rasch modelinde, modeldeki her bir yüzey için, ayırma oranı ve güvenirlik indeksi olmak üzere iki farklı güvenirlik istatistiği elde edilmektedir. Bu iki güvenirlik istatistiği farklı metriklerde rapor edilmesine rağmen; her iki istatistik de aynı bilgilerden hesaplanmakta ve belirli bir yüzey için benzer sonuçlara yol açmaktadır. Hem ayırma oranı hem de güvenirlik indeksi, herhangi bir

yüzeyin bileşenlerinin birbirinden hangi güvenirlikte ayrıldığını göstermektedir. Her bir yüzey için güvenirlik indeksi 0 ile 1 arasında değişmektedir. Ayırma oranı ise, 1 ile ∞ aralığında yer almaktadır (Sudweeks, Reeveb ve Bradshawc, 2004). Güvenirlik indeksi ile ayırma oranının anlamlılığını gösteren Ki Kare testi sonuçları incelenerek herhangi bir yüzeyin bileşenleri arasında anlamlı fark olup olmadığı belirlenebilmektedir. Örneğin, puanlamadaki katılık/cömertlikleri açısından puanlayıcılar arasında ya da yetenek düzeyleri açısından bireyler arasında anlamlı fark bulunup bulunmadığı, puanlayıcı ve birey yüzeyleri için hesaplanan Ki Kare değerleri ile tespit edilebilmektedir (Weigle, 1998).

Güvenirlik indeksi ve ayırma oranı çok yüzeyli Rasch analizindeki tüm yüzeyler için elde edilmesine karşın; bu iki istatistiğin yorumlanması her bir yüzey için farklılık göstermektedir. Birey ve madde yüzeyleri için, güvenirlik indeksinin ve ayırma oranının yüksek olması arzu edilen bir durumdur. Birey yüzeyine ait güvenirlik indeksinin yüksek olması yetenek düzeyleri farklı olan bireylerin birbirlerinden ayırt edilebildiği anlamına gelmektedir (Sudweeks, Reeveb ve Bradshawc, 2004). Birey yüzeyi için güvenirlik indeksi klasik test kuramındaki Cronbach Alpha güvenirlik katsayısına benzer şekilde yorumlanmaktadır. Güvenirlik indeksi, gerçek varyansın gözlenen varyansa oranını temsil etmektedir. Güvenirlik indeksinin 1'e yaklaşması yüksek düzeyde güvenirliğe işaret etmektedir. Birey yüzeyi için hesaplanan güvenirlik indeksi ile ayırma oranının düşük olması ise puanlamalara ranj sınırlaması etkisinin karıştığı anlamına gelmektedir. Bu nedenle, birey yüzeyi için tespit edilen düşük güvenirlik istatistikleri, puanlayıcıların bireyleri birbirinden ayırt etmede başarısız olduğu şeklinde yorumlanmaktadır. Madde yüzeyi için güvenirlik indeksi ve ayırma oranının yüksek olması ise kavramsal olarak farklı yapıları ölçen maddelerin birbirinden ayırt edilebildiğini göstermektedir. Madde yüzeyi için tespit edilen güvenirlik indeksinin ve ayırma oranının düşük olması ise, farklı özellikleri ölçen maddelerin birbirinden ayırt edilemediğini yansıtmaktadır. Dolayısıyla, ölçme işlemine halo etkisinin karışmış olabileceğine işaret etmektedir (Kramer vd., 2009).

Birey ve madde yüzeyi dışındaki diğer yüzeyler için güvenirlik indeksinin ve ayırma oranının düşük olması istenmektedir. Çünkü birey ve madde yüzeyi dışındaki yüzeylerin bileşenlerine ait değişkenlikler puanlamadaki varyans ile ilgisiz yapıları temsil etmektedir. Örneğin; puanlayıcı yüzeyine ait güvenirlik istatistiklerinin yüksek olması, öğrencilerin yetenek düzeylerine ilişkin kestirimlere puanlayıcı kaynaklı

faktörlerin karıştığını ve farklı puanlayıcılar tarafından yapılan puanlamalar arasında farklılıklar bulunduğunu yansıtmaktadır. Öte yandan, puanlayıcı yüzeyi için güvenirlik istatistiklerinin düşük olması, bu yüzeyin bileşenleri arasındaki tutarlığın göstergesi olarak kabul edilmektedir. Bir başka ifadeyle, puanlayıcı yüzeyi için hesaplanan güvenirlik, puanlayıcılar arasındaki güvenilir benzerliği değil; güvenilir farklılığı göstermektedir (Haiyang, 2010). Buna göre, puanlayıcı yüzeyi için elde edilen güvenirlik puanlayıcılar arası uyuma yönelik bir ölçü olmayıp, puanlayıcıların katılık ve cömertlikleri arasındaki farka ilişkin bir ölçüdür. Puanlayıcı yüzeyine ait güvenirlik, puanlayıcıların değerlendirdikleri bireyler için yaptıkları sıralama üzerinden değil (göreli uyum); bu bireylere verdikleri puanların gerçek değeri üzerinden (mutlak uyum) hesaplanmaktadır. (Sudweeks, Reeve ve Bradshaw, 2004).

FACET programından elde edilen istatistiklerden bir diğeri de yanlılık analizleridir. Yanlılık analizi, çok yüzeyli Rasch modeline dâhil edilen yüzeyler arasında sistematik bir etkileşim olup olmadığı hakkında bilgi vermektedir (Lumley ve McNamara, 1995; Mulqueen, Baker ve Dismukes, 2000). Analizde belirtilen yüzeyler arasındaki her bir etkileşim, logit ölçeğinde bir yanlılık puanı vermektedir. Bu puanın anlamlılığı standart *z* puanı (ya da *t* değerleri) ile belirlenmektedir. Mutlak değerce 2'ye eşit ya da 2'den yüksek *z* değerleri, yüzeyler arasındaki anlamlı etkileşimi göstermektedir (Sudweeks, Reeve ve Bradshaw, 2004).

2.2. İLGİLİ ARAŞTIRMALAR

İlgili araştırmalar, üç başlık altında sunulmuştur. Birinci başlık, standart rubrikler ile ilgili araştırmalardan oluşmuştur. Bu başlık altında sıralanan çalışmaların orijinalinde standart rubrik ifadesi geçmemektedir. Bununla birlikte, söz konusu araştırmalarda kullanılan rubrikler herhangi bir taksonomi referans alınmadan geliştirildiğinden bu araştırmalarda kullanılan rubriklerin standart rubrik olduğuna karar verilmiştir. Dolayısıyla, araştırmalar hakkında bilgi verilirken çalışmalardaki dereceli puanlama anahtarı ya da rubrik ifadelerinin yerine standart rubrik kavramı kullanılmıştır. İkinci başlıkta SOLO taksonomisi ile ilgili araştırmalar incelenmiştir. Üçüncü başlıkta ise, puanlayıcı etkilerinin çok yüzeyli Rasch modeli ile incelendiği çalışmalar özetlenmiştir. Araştırma konusu, kültürel özelliklerden etkilenmeye açık bir değişken olmadığından, ilgili araştırmalar sunulurken yurt içinde yapılan araştırmalar ve yurt dışında yapılan araştırmalar gibi bir sınıflandırmaya gerek duyulmamıştır.

2.2.1. Standart Rubrikler ile İlgili Araştırmalar

Kan (2007) tarafından yapılan çalışma ile *i)* rubrik tanımından hareketle onun doğasını ve ilkelerini ortaya çıkarmak, *ii)* rubriklere ilişkin uygulamalı tanımlar yapmak, *iii)* rubriklerin nitelik ve özelliklerini literatür ışığında tartışmak, *iv)* öğrenci kazanımlarını değerlendirmeye dönük rubrikleri geliştirmeyi ve yapılandırmayı amaçlayan eğitim literatürüne katkıda bulunmak, *v)* rubrik geliştirmek isteyen araştırmacı ve öğretmenlere tutarlı ve başarılı bir rubrik geliştirmek için yardımcı olmak amaçlanmıştır. Çalışmanın amaçlarına ulaşmak için öncelikle rubriklerin ne olduğu, türleri ve özellikleri aydınlatılmıştır. Daha sonra tutarlı ve başarılı bir rubrik geliştirebilmek için açıklayıcı örnekler ve şablonlarla birlikte uygulamalar, göz önünde bulundurulması gereken hususlar ve önerilere yer verilmiştir. Araştırma kapsamında örnek olarak sunulan rubrikler, standart rubrik şeklinde hazırlanmış ve rubriklerin geliştirilmesinde herhangi bir eğitim taksonomisi temele alınmamıştır.

Kasap (2008) tarafından yapılan araştırmada, puanlama anahtarı ile standart rubriklerden elde edilen puanların karşılaştırılması amaçlanmıştır. Araştırmaya genel matematik dersini alan 195 üniversite öğrencisi ile iki puanlayıcı dâhil edilmiştir. Öğrencilere açık uçlu 10 sorudan oluşan matematik başarı testi uygulanmıştır. Ardından, sınav kâğıtları puanlayıcılar tarafından önce puanlama anahtarı ve sonrasında standart rubrikler kullanılarak puanlanmıştır. Çalışmada ayrıca, öğrenciler standart rubrikler kullanarak öz değerlendirme ve akran değerlendirme yapmışlardır. Araştırma sonuçlarına göre; puanlayıcıların puanlama anahtarı ve standart rubrik kullanarak yaptıkları puanlamalar arasında anlamlı fark bulunduğu belirlenmiştir. Öğrencilerin standart rubrikleri kullanarak yaptıkları öz değerlendirme ve akran değerlendirme ile puanlayıcıların puanlama anahtarını kullanarak yaptıkları puanlamalar arasında anlamlı fark olduğu saptanmıştır. Yine, puanlayıcıların standart rubrikleri kullanarak yaptıkları puanlamalar ile öğrencilerin bu rubrikleri kullanarak yaptıkları öz ve akran değerlendirmeler arasında anlamlı fark bulunduğu tespit edilmiştir. Hem puanlama anahtarı hem de standart rubrik kullanılarak yapılan puanlamalar için iki puanlayıcı tarafından yapılan puanlamalar arasında oldukça yüksek korelasyonlar bulunmuştur. Puanlayıcıların standart rubrikler kullanarak yaptıkları puanlamaların iç tutarlılığının puanlama anahtarı kullanarak yaptıkları puanlamaların iç tutarlılığına göre daha yüksek olduğu sonucuna ulaşılmıştır.

Atmaz'ın (2009) yaptığı çalışmada, öğrencilerin grafik yorumlama becerilerini ölçmeye yönelik açık uçlu maddeler standart rubrik kullanılarak

puanlanmış ve yapılan puanlamaların güvenirliği incelenmiştir. Araştırmada öğrencilerin grafik yorumlama becerileri açık uçlu dört maddeden oluşan bir test yardımıyla ölçülmüştür. Öğrencilerin açık uçlu maddelere verdikleri yanıtların puanlanmasında standart rubriklerden yararlanılmıştır. Araştırma, sekizinci sınıfta öğrenim gören 50 öğrenci ve 17 matematik öğretmeninin katılımıyla gerçekleşmiştir. Açık uçlu sorular öğrenciler tarafından cevaplandıktan sonra, öğretmenler öğrencilerin yanıtlarını puanlama yönergesi kullanmadan puanlamışlardır. Bellek etkisini önlemek için iki hafta bekledikten sonra, öğretmenler puanlama yönergesi kullanarak aynı kâğıtları tekrar puanlamıştır. Zamanın puanlamalar üzerindeki etkisini ölçmek için aynı uygulama 10 hafta sonra tekrarlanmıştır. Bu şekildeki bir uygulamayla, aynı öğretmenlerin farklı zamanlarda verdikleri puanlar arasında anlamlı bir ilişki ve farklılık olup olmadığı belirlenmeye çalışılmıştır. Araştırmada, aynı öğretmenlerin farklı zamanlarda yaptıkları puanlamalar arasında pozitif yönde, oldukça yüksek ve anlamlı ilişkiler bulunmuştur. Fakat öğretmenlerin farklı zamanlarda verdikleri puanların ortalamaları arasında anlamlı farklılıklar olduğu belirlenmiştir. Yine, farklı öğretmenlerin puanları arasında pozitif yönde, oldukça yüksek ve anlamlı ilişkiler bulunduğu tespit edilmiştir. Bununla birlikte; farklı öğretmenlerin verdikleri puanlamalar arasında anlamlı fark olduğu saptanmıştır.

Ömür ve Erkuş (2013) tarafından yapılan araştırmada; genel izlenimle, standart rubrikle ve Thurstone ikili karşılaştırmalar yöntemiyle puanlanan kompozisyonlardan elde edilen verilere dayanarak, üç farklı puanlama yönteminin karşılaştırılması amaçlanmıştır. Yargıcıların kompozisyonları puanlamaları sonucunda, hangi puanlama yöntemi kullanılırsa kullanılsın kompozisyon sıralamalarında bir değişikliğin olmadığı saptanmıştır. Farklı yargıcıların beş ayrı kompozisyonu, genel izlenimle ve standart rubrik ile puanlanması sonucunda elde edilen verilere ilişkin puanlayıcı güvenirliğin belirlenmesinde G katsayısından yararlanılmış ve hesaplanan katsayılar yüksek bulunmuştur.

Özmen Hızarcıoglu (2013) tarafından yapılan çalışmada, problem çözme sürecinde standart rubrik kullanımının puanlayıcı uyumuna etkisi araştırılmıştır. Bu doğrultuda, puanlayıcılar problem çözme becerisini standart rubrik kullanarak ve kullanmadan puanlamış ve puanlayıcılar arasındaki uyum karşılaştırılmıştır. Araştırma, altıncı sınıfta okuyan 27 öğrenci ve 15 matematik öğretmeninin katılımıyla gerçekleşmiştir. Araştırmada veri toplama aracı olarak; öğrencilerin problem çözme becerilerini ölçmeye yönelik açık uçlu sorulardan oluşan bir test, bu

soruları puanlamak üzere geliştirilen standart rubrikler ve puanlayıcıların rubrikler hakkındaki düşüncelerini belirlemeye yönelik bir anket kullanılmıştır. Öğretmenler, öğrencilerin problemlere verdikleri yanıtları öncelikle rubrik kullanmadan puanlamışlardır. İki hafta sonra, aynı yanıtları standart rubrik kullanarak değerlendirmişlerdir. Puanlamaların ardından, öğretmenlerin standart rubrik kullanımına ilişkin görüşlerini belirlemek amacıyla, açık uçlu dört maddeden oluşan anket öğretmenlere uygulanmıştır. Öğretmenlerin öğrenci yanıtlarına verdikleri puanlar arasındaki uyum Kendall'ın w testi ile incelenmiştir. Öğretmenlerin rubrik kullanarak yaptıkları puanlamalar arasındaki uyumun yüksek olduğu gözlenmiştir. Rubrik kullanmadan verdikleri puanlar arasındaki uyum ise, düşük bulunmuştur. Öğretmenlerin standart rubrikler hakkındaki açıklamalarına göre; rubriklerin yapısı ve amacı hakkında fikir sahibi oldukları ancak hazırlama, uygulama ve çeşitleri hakkında yeterli bilgiye sahip olmadıkları gözlenmiştir. Öğretmenler standart rubriklerin objektif puanlamaya olanak tanıdığını, öğrencinin güçlü ve zayıf yönlerini görmesine yardımcı olduğunu, öğrenci başarısını olumlu yönde etkilediğini ifade etmişlerdir. Ancak rubrik hazırlamanın zor olduğunu ve uzun zaman aldığını bildirmişlerdir.

Parlak ve Doğan (2014) tarafından yapılan araştırmada, mesleki bilgi gerektiren uygulamalı bir derste, puanlama anahtarı ve standart rubrik kullanılarak elde edilen puanların uyum düzeyleri karşılaştırmalı olarak incelenmiştir. Araştırma, bir meslek lisesinin elektrik-elektronik bölümündeki 70 öğrenci ve altı öğretmenin katılımı ile gerçekleştirilmiştir. Araştırmadan elde edilen bulgulara göre; puanlama anahtarı ve rubrik kullanılarak elde edilen puanlar arasında göreli bir uyum olduğu, ancak mutlak bir uyum bulunmadığı belirlenmiştir. Araştırmada ayrıca, puanlama anahtarı ve rubrik kullanılarak hesaplanan puanlar ile öğrencilerin birinci dönem notları arasındaki korelasyona bakılmıştır. Hesaplanan korelasyon katsayıları karşılaştırıldığında, rubrik kullanılarak elde edilen puanların öğrencilerin karne notlarıyla ilişkisinin, puanlama anahtarıyla verilen puanların öğrencilerin karne notlarıyla ilişkisine göre daha yüksek olduğu tespit edilmiştir. Bununla birlikte, iki korelasyon katsayısı arasındaki farkın istatistiksel açıdan anlamlı olmadığı belirlenmiştir. Regresyon analizi sonucunda, rubrik ile verilen puanların öğrencilerin ders notlarını yordama düzeyinin daha yüksek olduğu saptanmıştır.

Standart rubrikler ile ilgili araştırmalara bakıldığında, bu rubriklerin puanlayıcı içi ve puanlayıcılar arası güvenirliğe etkisinin değişik puanlama

yöntemleri (genel izlenimle puanlama ve puanlama anahtarı ile puanlama gibi) ile karşılaştırılarak test edildiği görülmektedir. Ancak, standart rubriklerin puanlayıcı güvenirliğine etkisinin farklı eğitim taksonomilerinin temele alındığı rubrikler ile karşılaştırılarak sınandığı bir çalışmaya rastlanmamıştır.

2.2.2. SOLO Taksonomisi ile İlgili Araştırmalar

Hattie ve Purdie (1994) tarafından yapılan araştırmada, değerlendirme sorularının bilişsel düzeylerinin tespitinde SOLO ve Bloom taksonomisinin güvenirlikleri karşılaştırmalı olarak incelenmiştir. Çalışma kapsamında, öğretmenlere çoktan seçmeli 19 soru yöneltilmiş ve bu soruların karşılık geldiği bilişsel düzeyleri belirlemeleri istenmiştir. Araştırmanın katılımcılarını oluşturan 30 öğretmenin yarısı soruların karşılık geldiği bilişsel düzeyleri tespit etmede Bloom taksonomisinden yararlanırken; diğer yarısı soruların hangi bilişsel düzeye denk geldiğine karar vermek için SOLO taksonomisini esas almıştır. Araştırmadan elde edilen bulgular; Bloom taksonomisine göre yapılan sınıflandırmada öğretmenler arasındaki mutlak uyumun %40, öğretmenler arasında en fazla bir düzey fark bulunan sınıflandırmaların oranının ise %75 olduğunu göstermiştir. Diğer taraftan, SOLO taksonomisine göre yapılan sınıflandırmada öğretmenler arasındaki mutlak uyum %60, öğretmenler arasında en fazla bir düzey fark bulunan sınıflandırmaların oranı ise %96 olarak bulunmuştur.

Burnett (1999) tarafından yapılan araştırmada, danışmanlık hizmetinden kazanılan öğrenmelerin yapısını değerlendirmede kullanılabilecek bir model olarak SOLO taksonomisinin etkililiği araştırılmıştır. Araştırmaya son birkaç ay içerisinde kişisel problemlerden, ikili ilişkilerdeki sorunlara kadar farklı nedenlerle danışmanlık hizmeti alan 35 birey dâhil edilmiştir. Bu bireylerden, aldıkları danışmanlık hizmeti sayesinde öğrendiklerini ve danışmanlık hizmetinin kendilerine sağladığı yararları arkadaşlarına anlatan bir mektup yazmaları istenmiştir. Yazılan mektuplar SOLO taksonomisi kullanılarak sınıflandırılmış ve analiz edilmiştir. Araştırmada, SOLO taksonomisi beş düzeyli orijinal yapıdan farklı olarak sekiz düzeyli olarak ele alınmıştır. Yazılan mektuplar iki farklı puanlayıcı tarafından değerlendirilmiş ve puanlayıcılar arası güvenirlik %85 olarak hesaplanmıştır. Puanlayıcıların farklı düzeylere atadıkları yanıtlar incelendiğinde, cevabın atandığı temel düzeyin aynı (çok yönlü yapı gibi), ancak atamanın yapıldığı alt düzeyin farklı (çok yönlü yapı-düşük, çok yönlü yapı yüksek gibi) olduğu belirlenmiştir.

Leung (2000) tarafından yapılan çalışmada, Teknoloji ve Tasarım dersi kapsamında öğrencilerin tasarım yeteneklerini incelemek için SOLO taksonomisinin etkililiği araştırılmıştır. Çalışmada, SOLO taksonomisine dayalı olarak yeni bir madde formatı ve değerlendirmede kullanılacak ölçütlere ilişkin yeni bir çerçeve tasarlanmıştır. Araştırma 79 öğrencinin katılımıyla gerçekleşmiştir. Öğrencilerin tasarım performanslarını ölçmek için geleneksel maddeler ve SOLO taksonomisine uygun açık uçlu sorular hazırlanmıştır. Öğrencilerin geleneksel ve SOLO taksonomisine uygun bir biçimde hazırlanan maddelere verdikleri yanıtlar, hem geleneksel yöntemle hem de SOLO taksonomisine dayalı olarak iki puanlayıcı tarafından değerlendirilmiştir. Geleneksel yöntemle yapılan puanlamada öğrencilerin verdikleri yanıtlardaki nicel özelliklere odaklanılmıştır. Örneğin; öğrencinin bir soruya yönelik olarak kaç farklı tasarım önerdiği belirlenmiş ve bu doğrultuda puanlama yapılmıştır. SOLO taksonomisine göre yapılan puanlamalarda ise, nicel özelliklerin yanı sıra cevabın niteliği de göz önünde bulundurulmuştur. Söz gelimi, öğrencinin herhangi bir soruya yönelik olarak kaç farklı tasarım önerdiğiyle birlikte, önerilen bu tasarımların uygulanabilirliği de dikkate alınmıştır. Araştırmadan elde edilen bulgulara göre, SOLO taksonomisi doğrultusunda hazırlanan açık uçlu soruların, geleneksel olarak hazırlanan maddelere kıyasla teknoloji tasarım dersinin gerektirdiği yaratıcı düşünme gibi üst düzey düşünme becerilerinin ölçülmesi için daha uygun olduğu belirlenmiştir. Araştırmadan elde edilen bulgular, SOLO taksonomisine dayalı olarak yapılan değerlendirmelerin teknoloji ve tasarım dersinin amaçları ile örtüştüğünü göstermiştir.

Chan vd. (2002) tarafından yapılan çalışmada, bilişsel öğrenme çıktılarının değerlendirilmesinde, farklı eğitim taksonomilerinin kullanımı incelenmiştir. Araştırmada; SOLO taksonomisi, Bloom Taksonomisi ve yansıtıcı düşünme modeli olmak üzere üç farklı model kullanılmış ve bu taksonomilerin uygulanabilirlikleri test edilmiştir. Araştırma kapsamında, *i)* SOLO taksonomisine alt düzeyler eklenmesinin puanlayıcılar arası güvenirliği arttırıp arttırmadığı, *ii)* SOLO taksonomisinin farkı dersler ve farklı konulara ilişkin öğrenme çıktılarının değerlendirilmesi için uygun olup olmadığı, *iii)* farklı eğitim taksonomilerinin çapraz doğrulamasının değerlendirme işlemindeki güvenirliği arttırıp artırmadığı, *iv)* öğrencilerin eleştirel düşünme becerilerinin değerlendirilmesinde yansıtıcı düşünme modelinin Bloom ve SOLO taksonomisine kıyasla daha uygun olup olmadığı ve *v)* SOLO taksonomisinin Bloom taksonomisi ve yansıtıcı düşünme modeline göre daha

kullanışlı bir model olup olmadığı sorularına yanıt aranmıştır. Araştırma, farklı performans görevleri içeren iki ayrı çalışma şeklinde yürütülmüştür. Birinci çalışmadan elde edilen bulgulara göre, puanlayıcılar arası güvenirlik; SOLO taksonomisi kullanılarak yapılan puanlama için *r*=.60, Bloom taksonomisi kullanılarak yapılan puanlama için *r*=.93 ve yansıtıcı düşünme modeli kullanılarak yapılan puanlama için *r*=.87 şeklindedir. Üç taksonomiye göre yapılan puanlamalar arasındaki korelasyon incelendiğinde; SOLO taksonomisi ile yansıtıcı düşünme modeli arasındaki korelasyonun .75, SOLO taksonomisi ile Bloom taksonomisi arasındaki korelasyonun .74 ve Bloom taksonomisi ile yansıtıcı düşünme modeli arasındaki korelasyonun .84 olduğu saptanmıştır. İkinci çalışmadan elde edilen bulgulara göre, puanlayıcılar arası güvenirlik; SOLO taksonomisi kullanılarak yapılan puanlamalar için .66 ve Bloom taksonomisi kullanılarak yapılan puanlamalar için .68 olarak hesaplanmıştır. Yansıtıcı düşünme modeli kullanılarak yapılan değerlendirmede ise, iki puanlayıcı arasındaki korelasyon istatistiksel olarak anlamlı bulunmamıştır. Üç farklı taksonomiye göre yapılan puanlamalar arasındaki korelasyon incelendiğinde, SOLO taksonomisi ile yansıtıcı düşünme modeli arasındaki korelasyon *r*=.09 olarak hesaplanmıştır. Bloom taksonomisine göre yapılan puanlama ile diğer iki taksonomiye göre yapılan puanlama arasında ise, manidar bir korelasyonun bulunmadığı belirlenmiştir. SOLO taksonomisi; birinci ve ikinci çalışmada farklı konular ve dersler için kullanılmasına rağmen, her iki çalışmada da kabul edilebilir düzeyde güvenilir sonuçlar vermiştir. Buna göre, SOLO taksonomisinin farklı konu alanlarında ve derslerde kullanılabileceği belirlenmiştir. SOLO taksonomisinin beş düzeyli olarak kullanıldığı ikinci çalışmada hesaplanan puanlayıcılar korelasyon, dokuz düzeyli olarak kullanıldığı ikinci çalışma için hesaplanan puanlayıcılar arası korelasyondan daha yüksek bulunmuştur. Bununla birlikte, SOLO taksonomisinin dokuz düzeyli olarak kullanıldığı birinci çalışmada, beş düzeyli olarak kullanıldığı ikinci çalışmaya kıyasla; yetenek düzeyleri farklı olan öğrencilerin birbirinden daha iyi ayırt edilebildiği ve taksonomiye eklenen alt düzeylerin taksonomideki belirsizliği azaltarak puanlayıcılar arası güvenirliği arttırdığı sonucuna ulaşılmıştır. Araştırmada, öğrenme çıktılarının değerlendirmesinde farklı taksonomilerin çapraz doğrulaması yapıldığında, daha doğru değerlendirmeler elde edilebildiği tespit edilmiştir. Öğrencilerin eleştirel düşünme becerilerinin değerlendirilmesinde, yansıtıcı düşünme modelinin Bloom ve SOLO taksonomisine göre daha uygun olduğu saptanmıştır.

Kayani, Ajmal ve Rahman (2010) tarafından yapılan araştırmada, öğretmenlerin SOLO taksonomisine dayalı yeni bir sınav sistemi hakkındaki görüşleri incelenmiştir. Pakistan'da yürütülen araştırma 360 öğretmenin katılımıyla gerçekleşmiştir. Çalışmanın verileri, öğretmenlerin SOLO taksonomisine dayalı sınavlar hakkındaki düşüncelerini belirlemeye yönelik olarak araştırmacıların geliştirdiği ölçek ile toplanmıştır. Araştırmadan elde edilen bulgular, öğretmenlerin SOLO taksonomisine dayalı yeni sınav sistemini; sınavlarda geçerlik ve güvenirliği arttıran, puanlama işleminin daha kolay bir biçimde gerçekleştirmesini sağlayan, öğrencilerin yaratıcı düşünme, okuma, yazma ve anlama becerilerinin gelişimini destekleyen bir model olarak algıladığını ortaya koymuştur.

Alsaadi (2011) tarafından yapılan araştırmada, İngiltere ve Katar'da uygulanmakta olan matematik öğretim programlarında yer alan kazanımlar SOLO taksonomisi açısından analiz edilmiştir. Analiz sonucunda her iki ülkenin matematik programında SOLO taksonomisinin tek yönlü, çok yönlü ve ilişkisel yapı düzeylerine karşılık gelen kazanımlar olduğu belirlenmiştir. İki ülkenin matematik programları arasındaki tek farkın; SOLO taksonomisinin ilişkisel yapı düzeyi ile ilgili olduğu saptanmıştır. İlişkisel yapı düzeyinin, İngiltere'de uygulanmakta olan matematik programında Katar matematik programına göre daha fazla vurgulandığı tespit edilmiştir.

Kiani (2011) tarafından yapılan araştırmada, okul yöneticilerinin ve öğretmenlerin SOLO taksonomine dayalı yeni bir sınav sistemi hakkındaki düşüncelerinin incelenmesi, öğrencilerin SOLO taksonomisine dayalı bir başarı testindeki performanslarının analiz edilmesi ve SOLO taksonomisine dayalı yeni sınav sisteminin güçlü ve zayıf yönlerinin belirlenmesi amaçlanmıştır. Pakistan'da yürütülen araştırma, 120 okul yöneticisi, 360 öğretmen ve beşinci sınıfa devam eden 450 öğrencinin katılımıyla gerçekleştirilmiştir. Araştırmada okul yöneticilerinin ve öğretmenlerin SOLO taksonomisine dayalı yeni sınav sistemi hakkındaki görüşlerini incelemek için beşli derecelendirmeye sahip Likert tipi iki farklı ölçek kullanılmıştır. Öğretmenlere uygulanan ölçekte 30 madde, okul yöneticilerine uygulanan ölçekte ise 28 madde yer almıştır. Öğrencilerin sınav performanslarına ilişkin veriler ise; biri eski sınav sistemine, diğeri SOLO taksonomisine dayalı yeni sınav sistemine uygun olarak geliştirilen iki farklı başarı testi ile toplanmıştır. Elde edilen bulgular, araştırmaya katılan okul yöneticileri ile öğretmenlerin yeni sınav sisteminden memnun olduğunu göstermiştir. Katılımcıların büyük çoğunluğunun, SOLO

taksonomisine dayalı yeni sınav sistemini ezberciliği azaltan, değerlendirmelerin güvenirliğini arttıran, öğrencilerin öğrenmesini, yaratıcılığını, okuma ve yazma becerisini destekleyen ve kolay uygulanabilen bir model olarak gördüğü belirlenmiştir. Araştırmada ayrıca, öğrencilerin SOLO taksonomisine dayalı olarak geliştirilen testteki başarılarının, eski sınav sistemine uygun olarak hazırlanan testteki başarılarına göre daha yüksek olduğu saptanmıştır.

Arı (2013) tarafından yapılan araştırmada, bilişsel alan sınıflamasına yönelik olarak ileri sürülen Yenilenmiş Bloom, SOLO, Fink ve Dettmer taksonomileri hakkında eğitim programları ve öğretim anabilim dalı öğretim elemanları ile taksonomileri geliştiren uzmanların görüşleri incelenmiştir. Araştırmanın verileri; görüşme ve araştırmacı tarafından geliştirilen ölçek yardımıyla toplanmıştır. Araştırmadan elde edilen bulgulara göre, akademisyenlerin değişik alanlarda farklı taksonomilerin kullanılması gerektiğini savunduğu, Fink ve Dettmer taksonomisinin diğer taksonomilere göre daha az tanındığı, Bloom'un orijinal taksonomisinin ve bu taksonominin revize edilmiş halinin birçok ülkede yaygın olarak kullanıldığı belirlenmiştir.

Yazıcı (2013) tarafından yapılan çalışmada, SOLO taksonomisine dayalı olarak hazırlanan rubriklerin puanlayıcı güvenirliği üzerindeki etkisi geleneksel puanlama anahtarları ile karşılaştırılarak incelenmiştir. Araştırma verilerinin toplanmasında Fizik Dersi Madde ve Özellikleri ünitesine yönelik olarak hazırlanan altı açık uçlu sorudan oluşan bir başarı testinden yararlanılmıştır. Geliştirilen başarı testi, 11. sınıfa devam eden 200 lise öğrencisine uygulanmıştır. Uygulamanın ardından öğrenci cevapları üç puanlayıcı tarafından önce geleneksel puanlama anahtarı, ardından SOLO taksonomisine dayalı rubrikler kullanılarak puanlanmıştır. Araştırmada puanlayıcı güvenirliğinin belirlenmesinde klasik test kuramına dayalı yöntemlerden biri olan puanlayıcılar arası korelasyon katsayısından yararlanılmıştır. Puanlayıcılar arası korelasyon katsayılarının SOLO taksonomine dayalı rubrikler ile yapılan puanlamalarda geleneksel puanlama anahtarı ile yapılan puanlamalara kıyasla daha yüksek olduğu saptanmıştır.

Sıralanan araştırmalar; çok sayıda farklı derste ve birçok farklı eğitim kademesinde öğrencilerin açık uçlu sorulara verdikleri cevapların puanlanmasında SOLO taksonomisine dayalı rubriklerden yararlanıldığını göstermektedir. Bununla birlikte, literatürde SOLO taksonomisine dayalı rubrikler kullanılarak yapılan puanlamalarda, puanlayıcı etkilerini belirlemeye yönelik sınırlı sayıda araştırma

olduğu belirlenmiştir. Bu araştırmaların bir kısmında; SOLO taksonomisi, Bloom taksonomisi ve yansıtıcı düşünme modeline dayalı rubrikler için hesaplanan puanlayıcı güvenirlikleri karşılaştırmalı olarak incelenmiştir. Araştırmaların diğer bir kısmında ise, SOLO taksonomisine dayalı rubrikler için hesaplanan puanlayıcı güvenirlikleri puanlayıcıların kendi hazırladıkları puanlama anahtarları ile karşılaştırılmıştır. Alanyazında, SOLO taksonomisine dayalı rubriklerin puanlayıcı güvenirliği üzerindeki etkisini standart rubrikler ile karşılaştırarak inceleyen bir araştırmaya ise rastlanmamıştır. Ayrıca, SOLO taksonomisine dayalı rubriklerin puanlayıcı güvenirliği üzerindeki etkisini inceleyen literatürdeki mevcut çalışmalarda, klasik test kuramına dayalı teknikler kullanılmıştır. Bu nedenle literatürdeki araştırmalar açık uçlu soruların puanlanmasında SOLO taksonomisine dayalı rubrik kullanımının halo etkisi, puanlayıcı yanlılığı ve merkeze yönelme gibi puanlayıcı etkileri ile birey ve madde güvenirliklerini nasıl etkilediği sorusuna cevap olamamaktadır.

2.2.3. Puanlayıcı Etkilerinin Çok Yüzeyli Rasch Modeli ile İncelendiği Araştırmalar

Engelhard (1992) tarafından yapılan araştırmada, yazma yeteneğinin ölçülmesinde kullanılabilecek bir model olarak çok yüzeyli Rasch modeli tanıtılmıştır. Çalışmada, yazma yeteneğinin değerlendirilmesi amacıyla eyalet çapında yapılan bir sınavdan tesadüfi olarak 1000 öğrenci seçilmiştir. Araştırmadan elde edilen bulgulara göre; puanlayıcı eğitimlerinden sonra bile, puanlayıcıların katılık ve cömertliklerinde önemli farklılıklar olduğu belirlenmiştir. Ayrıca yazma görevleri arasında küçük fakat istatistiksel olarak anlamlı farklılıklar saptanmıştır. Araştırmada, kompozisyonlar aracılığıyla yazma yeteneğinin değerlendirilmesine yönelik büyük ölçekli sınavlarda karşılaşılan ölçüm sorunlarını çözmek için çok yüzeyli Rasch modelinin umut verici bir yaklaşım olduğu ifade edilmiştir.

Engelhard'ın (1994) yaptığı araştırmada; merkeze yönelme etkisi, halo etkisi, ranj sınırlaması, puanlayıcı katılığı ve cömertliği gibi puanlayıcı hatalarının tanıtılması amaçlanmıştır. Çok yüzeyli Rasch modeline dayalı olarak, puanlamaların niteliğinin değerlendirilmesinde kullanılacak ölçütler sunulmuştur. Araştırmada tesadüfi olarak belirlenen 264 kompozisyon 15 puanlayıcı tarafından değerlendirilmiştir. Elde edilen bulgulara göre; değerlendirmedeki katılık ve cömertlikleri yönüyle puanlayıcılar arasında anlamlı fark olduğu saptanmıştır.

Puanlayıcılardan ikisi için halo etkisi belirlenmiş ve bu puanlayıcıların öğrencilerin kompozisyonlarını analitik olarak puanlamak yerine; holistik bir biçimde puanlama eğilimde olduğu sonucuna ulaşılmıştır. Puanlamaların yaklaşık %80'inin puanlama ölçeğinin orta iki kategorisinde yer aldığı belirlenmiş ve puanlamalarda merkeze yönelme etkisinin bulunduğu tespit edilmiştir. Ham puanlar incelendiğinde, ranj sınırlaması etkisinin puanlamalara karıştığı saptanmıştır. Öğrencilerin yazma yeteneğine ilişkin düzeltilmiş puanlarına karışan ranj sınırlaması etkisinin ise daha az olduğu ortaya çıkmıştır.

Weigle (1998) tarafından yapılan araştırmada, çok yüzeyli Rasch modeli kullanılarak, puanlayıcı eğitimlerinin puanlayıcı hatalarını kontrol altına alma konusundaki etkililiği incelenmiştir. Araştırma kapsamında, deneyimli ve deneyimsiz puanlayıcılardan oluşan bir grup puanlayıcının puanlama işlemindeki tutarlılıkları, katılık ve cömertlikleri puanlayıcı eğitimlerinden önce ve sonra incelenmiştir. Çalışmada, İngilizce kompozisyon yazma becerisini ölçmeye yönelik bir sınavda 60 öğrencinin performansı sekizi tecrübeli ve sekizi tecrübesiz olmak üzere toplam 16 puanlayıcı tarafından değerlendirilmiştir. Araştırmada, puanlayıcı eğitiminden önce deneyimsiz puanlayıcıların daha cömert davrandığı ve bu puanlayıcıların daha tutarsız puanlamalar yaptığı belirlenmiştir. Araştırmadan elde edilen bulgulara göre, deneyimli ve deneyimsiz puanlayıcılar arasındaki farkın puanlayıcı eğitimlerinden sonra azaldığı saptanmıştır. Puanlayıcı eğitiminin ardından birçok puanlayıcının daha tutarlı puanlamalar yaptığı tespit edilmiştir. Ancak, puanlayıcıların katılık ve cömertlikleri arasındaki farkın devam ettiği sonucuna ulaşılmıştır. Bu sonuçlardan hareketle, puanlayıcı eğitimlerinin puanlayıcı içi güvenirliği arttırmada puanlayıcılar arası güvenirliği arttırmaya kıyasla daha etkili olduğu ifade edilmiştir.

Congdon ve McQueen (2000) tarafından yapılan araştırmada, puanlayıcıların katılık ve cömertliklerinin puanlama sürecindeki değişimi incelenmiştir. Araştırmada 8285 ilköğretim öğrencisinin performansı 16 puanlayıcı tarafından değerlendirilmiştir. Her bir öğrencinin performansı iki bağımsız puanlayıcı tarafından puanlamıştır. Puanlama işlemi yedi gün sürmüş ve puanlayıcılar ilk gün değerlendirdikleri performansı, puanlama işleminin son gününde tekrar değerlendirmiştir. Gerek farklı günlerde yapılan puanlamalar ayrı ayrı incelendiğinde, gerekse de tüm puanlamalar bir arada ele alındığında puanlayıcılar arasında anlamlı fark olduğu belirlenmiştir. Puanlayıcıların katılık ve cömertliklerine ilişkin günlük tahminler karşılaştırıldığında, yalnızca tüm puanlama periyodunun

ortalaması için anlamlı fark tespit edilmiştir. Değerlendirmede görev alan 16 puanlayıcıdan 10'unun son gün yaptığı değerlendirmeler ile ilk gün yaptığı değerlendirmeler arasında puanlama katılığı ve cömertliği açısından anlamlı fark saptanmıştır.

Chi (2001) tarafından yapılan araştırmada, analitik ve holistik rubrik kullanılarak yapılan puanlamalar çok yüzeyli Rasch modeli ile incelenmiştir. Çalışmada öğrencilerin Kore'nin kültürel, tarihi ve coğrafi özellikleri hakkındaki bilgilerini değerlendirmek için öğrencilerden, Kore'nin Güney Merkez Bölgesi'ne yapılacak bir seyahat planlamaları istenmiştir. Araştırmaya 42 öğrenci dâhil edilmiştir. Öğrencilerin performansı dört puanlayıcı tarafından holistik ve analitik rubrik kullanılarak puanlanmıştır. Araştırmadan elde edilen bulgular, öğrenciler arasındaki göreli sıralamanın kullanılan rubrik türünden etkilenmediğini ortaya koymuştur. Ancak öğrencilerin yetenek düzeylerine ilişkin mutlak kararların kullanılan rubrik türüne göre farklılık gösterdiği saptanmıştır. Puanlayıcı katılığı ve cömertliği açısından, analitik rubrikler ile yapılan puanlamaların holistik rubrik kullanılarak yapılan puanlamalara göre daha tutarlı sonuçlar verdiği belirlenmiştir. Araştırmadan elde edilen bulguların, performans değerlendirmede kullanılacak puanlama yönteminin seçiminde etkili olabileceği ifade edilmiştir.

Myford ve Wolfe (2003) tarafından yapılan araştırmada, puanlayıcı etkilerinin belirlenmesinde ve ölçülmesinde kullanılan bir yöntem olan çok yüzeyli Rasch modelinin puanlayıcılara tanıtılması amaçlanmıştır. Araştırma ile halo etkisi, merkeze yönelme etkisi, yanlılık, puanlayıcı katılığı ve cömertliği gibi puanlayıcı etkilerinin belirlenmesinde FACET paket programının nasıl kullanılacağına yönelik bilgi verilmesi hedeflenmiştir. Araştırmanın birinci bölümünde, çok yüzeyli Rasch modeli, modelin arka planı ve içeriği tanıtılmıştır. Çalışmada, son çeyrek yüzyıl içerisinde araştırmacılar tarafından incelenen puanlayıcı etkilerini tanıtan bir katalog sunulmuştur. Böylelikle, puanlayıcı etkilerinin tarihsel süreç içerisinde nasıl kavramsallaştırıldığına dair okuyuculara bir bakış açısı kazandırılması amaçlanmıştır. Daha sonra, puanlayıcı etkilerinin performans değerlendirmeyi nasıl etkilediği, bu etkileri ölçmek için kullanılan parametrelere ilişkin ölçütler ve söz konusu etkileri minimum düzeyde tutmak için kullanılabilecek stratejiler açıklanmıştır. Araştırmanın ikinci bölümünde ise çok yüzeyli Rasch modelinin kavramsal ve matematiksel açıklaması verilmiş, araştırmacıların puanlayıcı etkilerini incelerken çok yüzeyli Rasch modelini nasıl kullanabileceğine odaklanılmıştır.

Turner (2003) tarafından yapılan araştırmada, öğrencilerin resim yeteneklerini ölçmek için hazırlanan portfolyo değerlendirme sisteminin incelenmesi amaçlanmıştır. Çalışmada, 12 portfolyo projesinin 15 öğrenci tarafından tamamlanmasıyla elde edilen verilerden yararlanılmıştır. Öğrenciler tarafından tamamlanan projeler 10 profesyonel resim tasarımcısı tarafından değerlendirilmiştir. Değerlendirmelerde dört boyuttan oluşan bir puanlama ölçeği kullanılmıştır. Her bir öğrencinin projesi yansız olarak seçilen üç puanlayıcı tarafından değerlendirilmiştir. Araştırmada, öğrencilerin yetenek düzeylerini, projenin ve projeyi oluşturan dört boyutun güçlük düzeyini, puanlayıcı katılığı ve cömertliğini analiz etmek için çok yüzeyli Rasch modelinden yararlanılmıştır. Ham puan ortalamaları esas alındığında, üç öğrencinin yetenek düzeyinin kestirimlerin üstünde, üç öğrencinin yetenek düzeyinin ise kestirimlerin altında olduğu tespit edilmiştir. Çok yüzeyli Rasch modeli, puanlayıcıların katılık ve cömertliklerini dikkate alarak öğrencilerin puanlarını düzeltmiş ve düzeltme işlemi sonucunda altı öğrencinin sıralaması değişmiştir. Puanlayıcı katılığı ve cömertliği için ayırma indeksi güvenirliği .93 ve puanlayıcılar arası tutarlılık %43.3 olarak bulunmuştur. Madde yüzeyine ilişkin ayırma indeksi güvenirliği .73 olarak bulunmuş ve puanlayıcıların ölçülen performansın farklı boyutlarını birbirinden ayırt etmede başarısız olduğu belirlenmiştir. Araştırmada ayrıca, projenin ve proje boyutlarının güçlük düzeyinin öğrencilerin yetenek düzeylerine ilişkin kestirimlerde etkili olmadığı saptanmıştır.

Atılgan (2004) tarafından yapılan çalışmada, çok yüzeyli Rasch modeli ve genellenebilirlik kuramı analizleri gerçek veriler kullanılarak karşılaştırılmıştır. Araştırmada 2002 ve 2003 yıllarında yapılan Müzik Öğretmenliği Özel Yetenek Seçme Sınavı verilerinden yararlanılmıştır. Yapılan analizlerde, *i)* tek ve çok değişkenli modellerin istatistikleri karşılaştırılmış, *ii)* genellenebilirlik kuramının alternatif karar çalışmalarının gerçek durumlarla tutarlılıkları incelenmiş ve *iii)* genellenebilirlik kuramı ile çok yüzeyli Rasch modeli istatistikleri karşılaştırılmıştır. Araştırmada, alt boyutlardan oluşan testlerde genellenebilirlik kuramının tek ve çok değişkenli modellerinin aynı ölçüme durumu için farklı sonuçlar ürettiği saptanmıştır. G ve Phi katsayılarının modelden etkilendiği belirlenmiştir. Yine genellenebilirlik kuramı analizlerine göre, alternatif karar çalışmalarıyla puanlayıcı sayılarına ilişkin farklı senaryolar karşısında elde edilen G ve Phi katsayılarının gerçek durumda kestirilen G ve Phi katsayılarından farklı olduğu tespit edilmiştir. Araştırmada hem genellenebilirlik kuramının hem de çok yüzeyli Rasch modelinin

oldukça kullanışlı sonuçlar ürettiği belirlenmiştir. Bununla birlikte, değişkenlik kaynakları için genellenebilirlik kuramı ve çok yüzeyli Rasch modeli ile kestirilen sonuçların kısmen tutarlı olduğu saptanmıştır.

Myford ve Wolfe (2004) tarafından yapılan çalışmada, performans değerlendirmeye karışan puanlayıcı etkilerinin belirlenmesinde, çok yüzeyli Rasch modelinin uygulanabilirliği incelenmiştir. Araştırmada, yanlılık, merkeze yönelme etkisi, puanlayıcı katılığı ve cömertliği olmak üzere üç puanlayıcı etkisi ele alınmış ve bu puanlayıcı etkilerinin göstergeleri olan parametreler tanımlanmıştır. Gelişmiş bir yerleştirme sınavı kapsamında, adaylar tarafından yazılan kompozisyonlar puanlayıcı eğitimi almış puanlayıcılar tarafından değerlendirilmiştir. Değerlendirme işlemine karışan puanlayıcı etkilerini açığa çıkarmak için söz konusu puanlayıcı etkilerini belirlemeye yönelik her bir parametre incelenmiştir. Elde edilen bulgular sonucunda, sıralanan puanlayıcı etkilerini sergilediğinden şüphelenilen puanlayıcılar belirlenmiş ve sözü geçen puanlayıcı etkilerinin değerlendirme işlemine göz ardı edilemeyecek büyüklükte karıştığı saptanmıştır.

Atılgan (2005b) tarafından yapılan çalışmada, İnönü Üniversitesi Eğitim Fakültesi Müzik Öğretmenliği Özel Yetenek Seçme Sınavları çok yüzeyli Rasch modeli ile analiz edilmiştir. Analiz kapsamında, bireylerin yetenek düzeyleri, puanlayıcıların katılık ve cömertlikleri, görevlerin güçlük düzeyleri, uygunluk istatistikleri ile *puanlayıcı×birey, görev×birey* ve *puanlayıcı×görev* etkileşimleri araştırılmıştır. Adayların performansları; deşifre, çalma ve söyleme olmak üzere üç boyutta değerlendirilmiştir. Değerlendirmeler müzik öğretimi alanından dört öğretim üyesi tarafından gerçekleştirilmiştir. Öğretim üyeleri öğrencilerin performanslarını birbirinden bağımsız olarak puanlamıştır. Elde edilen bulgulara göre; puanlayıcıların bütün bireyler için puanlamalarının birbirlerine yakın ancak manidar düzeyde farklı olduğu, sınava giren 10 birey ile sınav kapsamında uygulanan görevlerden birinin uygunluk istatistiklerinin kabul edilebilir aralığın dışında yer aldığı, *puanlayıcı×birey, görev×birey* ve *puanlayıcı×görev* etkileşimlerinde yanlılıkların bulunduğu belirlenmiştir.

Elder, Knoch, Barkhuizen ve von Randow (2005) tarafından yapılan çalışmada, puanlayıcılara analitik olarak puanladıkları yazma görevleriyle ilgili dönütler verilmiş, puanlayıcıların bu dönütlere ilişkin tutumları ve dönütlerin puanlayıcılar üzerindeki etkileri incelenmiştir. Araştırma kapsamında puanlayıcılar geniş ölçekli bir sınavda, üniversite öğrencilerinin yazma görevlerini

değerlendirmiştir. Çevrimiçi bir eğitim programına katılan sekiz puanlayıcı, 50 sınav kâğıdını birbirinden bağımsız olarak puanlamıştır. Araştırmada, puanlama ölçeğinin kategorileri ile puanlayıcılar arasında yanlılığa işaret eden anlamlı bir etkileşim olup olmadığını belirlemek, her bir puanlayıcının kendi puanlamaları arasındaki tutarlığı saptamak ve puanlayıcılar arasında diğer puanlayıcılara göre daha katı ya da daha cömert puanlamalar yapan puanlayıcılar bulunup bulunmadığını tespit etmek için çok yüzeyli Rasch modeli kullanılmıştır. Gerçekleştirilen puanlayıcı eğitimlerin ardından puanlayıcılar 60 sınav kâğıdını değerlendirmiştir. Puanlayıcılara, puanlama işlemi ile ilgili geri dönütler verilmiş ve puanlayıcılardan bu geri dönütleri dikkate almaları istenmiştir. Puanlayıcıların geri bildirimlere yönelik tutamlarını ortaya çıkarmak için nitel verilerden yararlanılmıştır. Nitel verilerin analizi sonucunda, puanlayıcıların geri dönütlere yönelik olumlu tutuma sahip olduğu belirlenmiştir. Geri dönütlerin puanlayıcıların kendi puanlama davranışlarına ilişkin farkındalığını arttırdığı saptanmıştır. Puanlayıcıların geri bildirimlere yönelik algıları arasında farklılıklar bulunduğu tespit edilmiştir. Puanlayıcıların, eğitimlerden önce ve sonra yaptıkları değerlendirmeler karşılaştırıldığında, puanlayıcıların birçoğunun puanlamasında değişiklik olduğu, puanlayıcılar arasındaki tutarlığın arttığı ve madde yanlılıklarının azaldığı belirlenmiştir. Araştırmada puanlayıcı eğitimlerinin, puanlayıcı güvenirliğinin arttırılması konusunda yarar sağladığı ancak testin ayırt ediciliğini olumsuz yönde etkilediği sonucuna ulaşılmıştır. Buna bağlı olarak, karmaşık bir puanlayıcı eğitim programının getirdiği maliyetin yararından daha fazla olabileceği ifade edilmiştir.

Sudweeks, Reeve ve Bradshaw (2005) tarafından yapılan araştırma, üniversitede öğrenim gören öğrencilerin kompozisyon yazma becerilerini değerlendirme sürecini geliştirmek amacıyla pilot bir çalışma olarak yürütülmüştür. Araştırmada, puanlamadaki hata kaynaklarını belirlemek, güvenilir kestirimler elde etmek ve değerlendirme sürecini geliştirecek öneriler getirmek için genellenebilirlik kuramı ve çok yüzeyli Rasch modelinden yararlanılmıştır. Araştırmada 497 üniversite öğrencisi tarafından yazılan kompozisyonlar incelenmiş ve bu kompozisyonlardan 48 tanesi puanlanmak üzere araştırmaya dâhil edilmiştir. Araştırmada, dokuz puanlayıcı, dokuzlu derecelendirmeye sahip holistik bir rubrik kullanarak 48 denemenin her birini iki defa puanlamışlardır. Puanlayıcıların her biri 24 öğrenci tarafından yazılan iki kompozisyonu puanlamıştır. Böylelikle "*öğrenci×görev×puanlayıcı×zaman*" şeklinde tam bir çapraz desen elde edilmiştir.

Bu çapraz desen sayesinde, puanlayıcılar tarafından iki farklı zamanda yapılan puanlamalar arasındaki tutarlılık, bir puanlayıcının farklı zamanlarda yaptığı puanlamalar arasındaki tutarlılık ve yazılan kompozisyonlar arasındaki tutarlılık incelenebilmiştir. Araştırma sonucunda, görev ve *birey×görev* etkileşiminden kaynaklanan varyansın yüksek; puanlayıcı ve puanlama zamanından kaynaklanan varyansın ise düşük olduğu saptanmıştır. Puanlamadaki değişkenliğin %22'sinin kaynağının açıklanamadığı belirlenmiştir. Araştırma kapsamında genellenebilirlik kuramı ile çok yüzeyli Rasch modelinin ortak özellikleri, farklılıkları, yararları ve sınırlılıkları tartışılmıştır.

Abu Kassim (2007) tarafından yapılan çalışmada, performans değerlendirmeye karışan puanlayıcı etkilerini belirlemek ve kontrol altına almak için çok yüzeyli Rasch modeli kullanılmıştır. Araştırmada Uluslararası Malezya İslam Üniversitesi Temel Araştırmalar Merkezi'nde görev yapan 34 İngilizce öğretmeni puanlayıcı olarak görev almıştır. Puanlayıcılar; Malezya Uluslararası İslam Üniversitesi tarafından yapılan seviye tespit sınavına giren öğrenciler tarafından yazılan paragrafları değerlendirmiştir. Araştırmada öğrenciler için elde edilen uyum istatistiklerinin kabul edilebilir sınırlar içerisinde yer aldığı belirlenmiştir. Puanlama işlemindeki katılık ve cömertlikleri açısından puanlayıcılar arasında anlamlı fark olduğu saptanmıştır. Puanlayıcılar arasındaki mutlak uyum oranı %29 olarak tespit edilmiştir. Puanlayıcı yüzeyine için hesaplanan uyum istatistikleri incelendiğinde, puanlayıcılardan üçü için uyum istatistiklerinin düşük olduğu ve bu puanlayıcılar tarafından yapılan değerlendirmelere ranj sınırlaması etkisinin karıştığı belirlenmiştir. Puanlayıcılardan beşinin ise uyum istatistiklerinin yüksek olduğu ve bu puanlayıcıların değerlendirmeler sırasında tutarsız davranarak benzer performanslara farklı puanlar verdiği sonucuna ulaşılmıştır.

Knoch, Read ve von Randow (2007) tarafından yapılan çalışmada, yüz yüze ve çevrimiçi olarak gerçekleştirilen puanlayıcı eğitimlerinin merkeze yönelme etkisi, halo etkisi, yanlılık, puanlayıcı katılığı ve cömertliği gibi puanlayıcı hatalarını azaltma konusundaki etkililiği karşılaştırmalı olarak araştırılmıştır. Araştırmada ayrıca, puanlayıcıların yüz yüze ve çevrimiçi olarak gerçekleştirilen puanlayıcı eğitimleri hakkındaki görüşleri incelenmiştir. Araştırmaya 16 puanlayıcı dâhil edilmiştir. Eğitim aşamasında, puanlayıcılar her bir grupta sekiz puanlayıcı olacak şekilde iki gruba ayrılmıştır. Puanlayıcı eğitimleri gruplardan birinde çevrimiçi, diğerinde ise yüz yüze gerçekleştirilmiştir. Araştırmadan elde edilen bulgular; hem

yüz yüze hem de çevrimiçi eğitimlerin puanlayıcı hatalarını azaltma konusunda etkili olduğunu göstermiştir. Bununla birlikte, puanlayıcı katılığı ve cömertliğini azaltma konusunda çevrimiçi eğitimlerin yüz yüze eğitimlerine göre nispeten daha etkili olduğu saptanmıştır. Halo etkisini azaltma açısından ise, yüz yüze puanlayıcı eğitimlerinin çevrimiçi puanlayıcı eğitimlerime kıyasla kısmen daha etkili olduğu belirlenmiştir. Nitel veriler analiz edildiğinde, puanlayıcıların çevrimiçi ve yüz yüze puanlayıcı eğitimlerden meydana gelen karma bir yöntemi tercih ettikleri sonucuna ulaşılmıştır.

Wiseman (2008) tarafından yapılan araştırmada, öğrencilerin yabancı dilde yazma yeteneklerini çok yüzeyli Rasch modeli ile incelenmiştir. Araştırmaya 39 öğrenci ve sekiz puanlayıcı dâhil edilmiştir. Öğrencilere hikâye ve bir konu hakkında okuyucuyu ikna etmeye esasına dayalı iki farklı yazma görevi verilmiştir. Öğrenciler, bir öykü ve okuyucuyu bir konu hakkında ikna etmeye dayalı bir yazı olmak üzere iki farklı kompozisyon yazmıştır. Elde edilen kompozisyonlar sekiz puanlayıcı tarafından analitik ve holistik rubrik kullanılarak puanlanmıştır. Analizlerde, öğrencilerin performansları üzerinde etkili olan; yazma yeteneği, puanlayıcı katılığı ve cömertliği, kompozisyonun konusu (hikâye/okuyucu bir konu hakkında ikna etmeye dayalı yazı) ve puanlama yöntemi (analitik/holistik) olmak üzere dört temel etki üzerinde durulmuştur. Çok yüzeyli Rasch analizinden elde edilen bulgulara göre, analitik ve holistik rubrik kullanılarak yapılan puanlamalarda öğrencilerin yazma yetenekleri açısından farklı düzeylere ayrıldığı belirlenmiştir. Puanlayıcıların öğrencilerin performanslarını farklı cömertlik/katılıkta puanladığı ve en katı puanlayıcının puanlama işlemi boyunca genel olarak katı davrandığı belirlenmiştir. Yazma görevi ile ilgili konuların eşit olmadığı saptanmış ve hikâye tarzında yazılan yazılarda yüksek puan almanın, bir konu hakkında okuyucuyu ikna etmek amacıyla yazılan yazılara kıyasla daha zor olduğu belirlenmiştir. Araştırmada puanlayıcıların karar alma davranışları, puanlama sırasında kaydettikleri ses kayıtlarından yararlanılarak incelenmiştir. Elde edilen bulgular; puanlayıcıların katılık/cömertliklerinin çeşitli davranışlarla ilişkili olduğunu göstermiştir. Puanlayıcının kendini izleme davranışı, analitik bir puanlama yerine holistik puanlama yapma eğilimde olması ya da puanlama ölçeğinin orta kategorilerini kullanmaya meyilli olması puanlamadaki katılığını azaltmaktadır. Yorgunluk gibi dışsal faktörlerin ve kötü el yazısının bazı puanlayıcıların daha katı puanlamalar yapmasına sebep olduğu belirlenmiştir. Puanlayıcıların geçmiş deneyimlerinin

puanlamaları üzerinde etkili olduğu tespit edilmiştir. Araştırmada ayrıca, puanlayıcıların analitik rubrik kullanarak yaptıkları değerlendirmelerde daha katı puanlamalar yaptığı sonucuna ulaşılmıştır.

Barkaoui (2008) tarafından yapılan çalışmada, puanlama yönteminin ve puanlayıcı tecrübesinin İngilizce kompozisyon yazma performansının değerlendirilmesi üzerindeki etkisi incelenmiştir. Araştırmaya 31'i deneyimsiz ve 29'u deneyimli olmak üzere toplam 60 puanlayıcı ile 180 öğrenci dâhil edilmiştir. Puanlayıcıların her biri tesadüfi olarak seçilen 24 kompozisyonu analitik ve holistik rubrik kullanarak iki defa puanlamıştır. Puanlayıcılar ile yapılan görüşmelerden ve puanlayıcıların değerlendirmeler sırasında kaydettikleri ses kayıtlarından yararlanılarak puanlayıcıların karar verme davranışları hakkında bilgi edinilmiştir. Değerlendirme sonuçları; puanlayıcı katılığı ve cömertliği ile puanlayıcı tutarlığını, analitik ve holistik rubrikler arasındaki ilişkiyi ortaya koymak için incelenmiştir. Araştırmadan elde edilen bulgulara göre; analitik ve holistik rubriklerin aynı yapıyı ölçtüğü belirlenmiştir. Yazma yetenekleri farklı olan öğrencileri birbirinden ayırt etme konusunda analitik rubriklerin holistik rubriklere göre daha etkili olduğu saptanmıştır. Holistik rubrik kullanılarak yapılan değerlendirmelerde puanlayıcılar arası, analitik rubrikler kullanılarak yapılan değerlendirmelerde ise, puanlayıcı içi güvenirliğin daha yüksek olduğu tespit edilmiştir. Analitik rubriklerin, özellikle deneyimsiz puanlayıcıların daha tutarlı puanlamalar yapmasına katkı sağladığı ve puanlayıcı içi güvenirliği arttırdığı belirlenmiştir. Deneyimsiz puanlayıcılarda, puanlayıcılar arası ve puanlayıcı içi değişkenliğin daha fazla olduğu, bu puanlayıcıların puanlama ölçeklerine daha fazla başvurduğu, metnin yorumlanmasına/düzenlenmesine daha fazla zaman ayırdığı ve daha çok yazma performansının belli bölümlerine odaklandığı saptanmıştır. Deneyimli puanlayıcıların puanlama ölçeğinin dışında yer alan ölçütlere daha fazla başvurma eğiliminde olduğu, metnin okunmasına ve değerlendirmesine daha fazla zaman ayırdığı, puanlama konusundaki özgüvenlerinin daha yüksek olduğu ve daha tutarlı puanlamalar yaptığı tespit edilmiştir. Puanlamada kullanılan rubrik türünün deneyimsiz puanlayıcıların iç tutarlılıklarında ve deneyimli puanlayıcıların katılık/cömertliklerinde önemli bir etkiye sahip olduğu saptanmıştır. Puanlama yönteminin puanlayıcı davranışları üzerindeki etkisinin deneyimsiz puanlayıcılarda daha belirgin olduğu tespit edilmiştir. Araştırmadan elde edilen bulgular genel olarak, İngilizce yazma performansının değerlendirilmesinde analitik rubriklerin

daha uygun olduğunu göstermiştir. Bununla birlikte, her iki rubrik türünün de farklı değerlendirme amaçları, değişik içerik türleri, farklı puanlayıcılar ve öğrenciler için uygun olabileceği ifade edilmiştir.

Güler (2008) tarafından yapılan araştırmada, matematik başarısının ölçülmesiyle elde edilen puanlar klasik test kuramı, çok yüzeyli Rasch modeli ve genellenebilirlik kuramına göre analiz edilmiştir. Üç kurama göre, puanların güvenirlikleri hesaplanmış ve elde edilen sonuçlar karşılaştırılmıştır. Araştırmada, TIMSS-1999'da yer alan açık uçlu sorulardan 24'ü 2007 yılı bahar döneminde 203 öğrenciye uygulanmıştır. Daha sonra öğrencilerin verdikleri cevaplar dört farklı puanlayıcı tarafından holistik rubrik kullanılarak değerlendirilmiştir. Ardından elde edilen puanların güvenirliği farklı kuramlara göre incelenmiştir. Klasik test kuramında, iç tutarlık güvenirliğinin belirlenmesinde Cronbach Alpha güvenirlik katsayısı kullanılmış, puanlayıcılar arası güvenirliğin tespitinde ise, Kendall Konkordans katsayısı, korelasyon katsayısı ve puanlamalar arasındaki farkın anlamlılığını gösteren F testinden yararlanmıştır. Genellenebilirlik kuramında, "*birey×görev×madde*" şeklinde tam bir çapraz desen kullanılarak, genellenebilirlik ve güvenirlik katsayıları incelenmiştir. Çok yüzeyli Rasch modeli ile yapılan analizlerde; birey, puanlayıcı ve madde yüzeyleri için ayrı ayrı güvenirlik katsayıları hesaplanmıştır. Klasik test kuramından elde edilen bulgular; matematik başarısının ölçülmesiyle elde edilen puanların iç tutarlığının .92, puanlayıcılar arası uyumu gösteren Kendall Konkordans katsayısının .52 olduğunu ve puanlayıcılar arasındaki korelasyon katsayılarının .90 ile .97 arasında değiştiğini göstermiştir. Ancak F testi sonuçları, puanlamalar arasında anlamlı farklılıkların olduğunu ortaya koymuştur. Genellenebilirlik kuramına göre, matematik başarısının ölçülmesiyle elde edilen puanların genellenebilirlik katsayısı .92 ve güvenirlik katsayısı .90 olarak hesaplanmıştır. Puanlayıcı değişkenlik kaynağının toplam varyansı açıklama oranı %2.1 olarak bulunmuştur. Çok yüzeyli Rasch modeline göre yapılan analizlerde, öğrenci yüzeyi için hesaplanan güvenirlik katsayı .95, puanlayıcı yüzeyi için hesaplanan güvenirlik katsayısı ise .99 olarak elde edilmiştir.

Schaefer (2008) tarafından yapılan çalışmada, anadili İngilizce olan puanlayıcılar tarafından değerlendirilen İngilizce kompozisyonlar yanlılık açısından incelenmiştir. Araştırmada görev alan 40 puanlayıcı, TOEFL İngilizce yazma testinden uyarlanan bir konu hakkında 40 üniversite öğrencisi tarafından yazılan kompozisyonları değerlendirmiştir. Puanlamalar; içerik, organizasyon, biçim,

açıklamaların kalitesi, dil kullanımı, mekanik ve akıcılık olmak üzere altı kategoriden oluşan analitik bir rubrik kullanılarak yapılmıştır. Elde edilen veriler çok yüzeyli Rasch modeline göre analiz edilmiştir. Rasch analizleri alt puanlayıcı gruplarının bazılarında, tekrarlanan yanlılık örüntüleri bulunduğunu göstermiştir. Puanlama ölçeğinin boyutları ile puanlayıcı etkileşimleri incelendiğinde, içerik ve organizasyon boyutlarını katı bir biçimde değerlendiren puanlayıcıların, dil kullanımı ve mekanik boyutlarını daha cömert bir biçimde değerlendirdikleri belirlenmiştir. Puanlayıcı ve öğrenci etkileşimleri incelendiğinde, puanlayıcıların yüksek yetenek düzeyindeki bireyleri puanlarken, düşük yetenek düzeyindeki bireylere kıyasla daha katı ya da daha cömert puanlamalar yaptığı saptanmıştır. Ayrıca bazı puanlayıcıların, yüksek yetenek düzeyindeki bireyleri beklenenden daha cömert, düşük yetenek düzeyindeki bireyleri ise beklenenden daha katı bir biçimde değerlendirdiği sonucuna ulaşılmıştır.

Johnson ve Lim (2009) tarafından yapılan araştırmada, yazma performansının değerlendirilmesinde puanlayıcıların anadili ile ilgili yanlılıkların incelenmesi amaçlanmıştır. Araştırmada, İngilizce sınavı kapsamında öğrencilerin yazdıkları kompozisyonlar anadili farklı olan 19 puanlayıcı tarafından değerlendirilmiştir. Ancak sınırlı sayıda puanlama yapan iki puanlayıcı veri setinden çıkarılmış ve analizler 17 puanlayıcı üzerinden gerçekleştirilmiştir. Değerlendirmede görev alan puanlayıcıların 13'ünün anadili İngilizce, diğer dördünün anadili ise İngilizce'den farklı bir dildir. Sınava giren öğrenciler 21 farklı dil grubuna ayrılmıştır. Araştırmadan elde edilen veriler çok yüzeyli Rasch modeline göre analiz edilmiştir. Rasch modeli ile her bir dil grubu için puanlayıcıların sergilediği yanlılık miktarının belirlenmesi, puanlayıcıların katılık/cömertliklerinin ortaya konulması ve söz konusu puanlayıcı etkilerinin kontrol altına alınması amaçlanmıştır. Araştırma sonuçları, tüm dil grupları ve bütün puanlayıcılar için yanlılığın minimum düzeyde olduğunu ve puanlamalarda anadil ile ilgili herhangi bir yanlılığın bulunmadığını göstermiştir.

Kim (2009) tarafından yapılan araştırmada, ana dili İngilizce olan ve ana dili İngilizce olmayan öğretmenlerin, öğrencilerin İngilizce sözel performansını değerlendirmedeki davranışları araştırılmıştır. Araştırmada ana dili İngilizce olan 12 öğretmen ile ana dili İngilizce olmayan 12 öğretmenin yaptığı puanlamalar; tutarlılık, puanlayıcı katılığı ve cömertliği ile değerlendirme ölçütleri açısından karşılaştırılmıştır. Araştırmadan edilen bulgular, ana dili İngilizce olan ve olmayan öğretmenlerin çoğunun kabul edilebilir düzeyde tutarlı puanlamalar yaptığını

göstermiştir. Hem anadili İngilizce olan öğretmenlerin bulunduğu grupta hem de anadili İngilizce olmayan öğretmenlerin bulunduğu grupta yalnızca bir ya da iki puanlayıcının tutarsız olduğu saptanmıştır. Puanlayıcıların, farklı görevler boyunca puanlamalarındaki katılıklarının benzer olduğu tespit edilmiştir. Bununla birlikte, öğretmenlerin değerlendirmede kullandıkları puanlama ölçütleri arasında önemli farklılıklar olduğu ortaya çıkmıştır. Araştırmadan elde edilen nitel veriler; telaffuz, özel gramer bilgilerinin kullanımı ve aktarılan bilgilerin doğruluğu açısından analiz edilmiştir. Analiz sonucunda, anadili İngilizce olan öğretmenlerin anadili İngilizce olmayan öğretmenlere kıyasla değerlendirmelerinde daha ayrıntılı ve özenli olduğu belirlenmiştir.

Matsuno (2009) tarafından yapılan çalışmada, öğretmenler tarafından yapılan değerlendirmeler ile öğrenciler tarafından yapılan öz değerlendirme ve akran değerlendirmenin karşılaştırılması amaçlanmıştır. Çalışmaya, 91 öğrenci ve dört öğretmen dâhil edilmiştir. Araştırmada, yüksek yetenek düzeyindeki öğrenciler başta olmak üzere, öğrencilerin öz değerlendirme yaparken tahmin edilenden daha düşük puanlama yaptığı belirlenmiştir. En cömert puanlamaların akran değerlendirme sırasında yapıldığı saptanmıştır. Bununla birlikte, öğrencilerin akranları değerlendirirken yüksek yetenek düzeyindeki bireylere daha katı, düşük yetenek düzeyindeki bireylere daha cömert davrandıkları ortaya çıkmıştır. Akran değerlendirme sırasında gözlenen bu eğilimin, öğrencilerin kendi performansından bağımsız olduğu belirlenmiş ve dolayısıyla yüksek yetenek düzeyindeki öğrencilerin daha katı; düşük yetenek düzeyindeki öğrencilerin ise daha cömert puanlamalar yaptığı hipotezi reddedilmiştir. Araştırmada ayrıca, akran değerlendirmeye ilişkin iç tutarlılığın öz değerlendirme ve öğretmenler tarafından yapılan değerlendirmelere kıyasla daha yüksek olduğu belirlenmiştir. Akran değerlendirmede yanlılığa işaret eden anlamlı etkileşimlerin öz değerlendirme ve öğretmenler tarafından yapılan değerlendirmeye göre daha az olduğu sonucuna ulaşılmıştır. Puanlayıcıların tutarlı puanlamalar yaptığı ancak puanlamaların yanlılık içerdiği saptanmıştır. Akran ve öz değerlendirme ile öğretmenler tarafından yapılan değerlendirmelerde, puanlayıcıların gramer özelliklerini puanlarken daha katı; yazım özelliklerini puanlarken daha cömert davranışlar sergilediği tespit edilmiştir. Analizler ayrıca öğretmenlerin; performansın yazım, format ve telaffuz boyutlarını diğer boyutlara göre daha farklı puanladığını göstermiştir. Araştırmada öz değerlendirmeden elde edilen sonuçların akran değerlendirme ve öğretmenler tarafından yapılan değerlendirmelere göre daha

farklı olduğu ve dolayısıyla öz değerlendirmenin formal değerlendirme süreçlerinin bir parçası olarak kullanımının sınırlı olduğu tespit edilmiştir. Akran değerlendirmenin iç tutarlılığının yüksek olduğu ve öğrencilerin akranlarını değerlendirirken kendi performanslarından etkilenmedikleri sonucuna ulaşılmıştır. Araştırmadan elde edilen sonuçlar, akran değerlendirmenin öğrencilerin yazma performanslarının değerlendirilmesinde yararlı olacağı şeklinde yorumlanmıştır.

Myford ve Wolfe (2009) tarafından yapılan araştırmada, puanlayıcılar tarafından yapılan değerlendirmelerin zaman içerisindeki değişimine yönelik bir çerçeve çizilmesi amaçlanmıştır. Araştırmada puanlayıcıların performanslarına ilişkin çeşitli parametreler sunulmuş ve bu parametrelerin operasyonel olarak nasıl kullanılacağı açıklanmıştır. Çalışmada, 2002 Yılı İngiliz Edebiyatı Yerleştirme Sınavı ile kompozisyon sınavına ait veriler kullanılmıştır. Elde edilen bulgulara göre, değerlendirme süreci ilerledikçe bazı puanlayıcıların ölçek kategorilerini kullanımlarında ve puanlamalarında anlamlı değişiklikler meydana geldiği tespit edilmiştir.

Haiyang (2010) tarafından yapılan araştırmada bilgi iletişim teknolojileri, ekonomi, mühendislik gibi farklı bölümlerden üniversite mezunu 56 öğrencinin bir İngilizce sınavından aldıkları puanlar klasik test kuramı ve çok yüzeyli Rasch modeline göre analiz edilmiştir. Öğrencilere uygulanan İngilizce testi; dinlediğini anlama, okuduğunu anlama, kelime bilgisi, çeviri ve yazma olmak üzere beş bölümden oluşmaktadır. Bunlardan, dinlediğini anlama, okuduğunu anlama ve kelime bilgisi bölümleri çoktan seçmeli maddelerden oluşmaktadır. Dolayısıyla öğrencilerin bu bölümlerden aldıkları puanlar objektif bir biçimde belirlenebilmektedir. Çeviri ve yazma bölümleri ise, objektif olarak puanlanamamakta; puanlayıcı yargılarına göre değerlendirilmektedir. Testin objektif olarak puanlanamayan bölümleri iki puanlayıcı tarafından holistik bir rubrik kullanılarak değerlendirilmiştir. Araştırmada klasik test kuramına göre yapılan analizlerde SPSS paket programından yararlanılmıştır. Klasik test kuramına göre gerçekleştirilen analizlerde; puanlayıcılar arası güvenirliğin belirlenmesinde Sperman sıra farkları korelasyonu, iç tutarlılığın ölçülmesinde ise Cronbach Alpha güvenirlik katsayısı kullanılmıştır. Testin objektif olarak puanlanamayan bölümleri ise çok yüzeyli Rasch modeline göre analiz edilmiştir. Analizlerde; öğrenciler, puanlayıcılar ve görevler olmak üzere üç yüzey yer almıştır. Araştırmada klasik test kuramına göre yapılan analizlerde, öğrencilerin İngilizce testinin objektif olarak

puanlanabilen bölümlerinden aldıkları notlar ile puanlayıcı yargılarına dayalı olarak puanlanan bölümlerinden aldıkları notlar arasındaki ilişkinin anlamlı olmadığı belirlenmiştir. Ayrıca testin objektif olarak puanlanabilen bölümlerindeki güvenirlik katsayısının yeterli düzeyde olduğu; öte yandan objektif olarak puanlanamayan bölümler için hesaplanan güvenirlik katsayılarının kabul edilebilir sınırların altında yer aldığı saptanmıştır. Çok yüzeyli Rasch analizinden elde edilen bulgular, iki puanlayıcının puanlamadaki katılık ve cömertlikleri arasında farklılıklar olduğunu, görevlerin güçlük düzeyleri açısından farklılık gösterdiğini, bazı öğrenciler ile görevler arasında yanlılığa işaret eden anlamlı etkileşimlerin bulunduğunu göstermiştir. Çok yüzeyli Rasch modeline göre; öğrenci yüzeyi için hesaplanan güvenirlik katsayısı .77, görev yüzeyi için hesaplanan güvenirlik katsayısı .97 ve puanlayıcı yüzeyi için hesaplanan güvenirlik katsayısı ise .91 olarak bulunmuştur.

Baştürk (2010) tarafından yapılan araştırmada, öğrencilerin bilimsel araştırma dersi ile ilgili kazanımları performansa dayalı olarak oluşturdukları ödevler aracılığıyla değerlendirilmiştir. Öğrenciler tarafından hazırlanan ödevler çok yüzeyli Rasch modeli kullanılarak incelenmiştir. Araştırma, Pamukkale Üniversitesi Eğitim Fakültesi Eğitim Bilimleri Bölümü Rehberlik ve Psikolojik Danışmanlık anabilim dalında "Bilimsel Araştırma Yöntemleri" dersine devam eden 20 öğrenci ile yürütülmüştür. Ayrıca tesadüfi olarak seçilen altı öğrenci araştırmada puanlayıcı olarak görev yapmıştır. Araştırmada; öğrenciler tarafından hazırlanan çalışmalar, puanlayıcılar ve puanlamada kullanılan ölçütler olmak üzere üç yüzey bulunmaktadır. Elde edilen bulgular incelendiğinde, öğrenciler tarafından yapılan çalışmalar arasında anlamlı fark bulunduğu belirlenmiştir. Öğrenciler tarafından hazırlanan çalışmaların yer aldığı yüzey için hesaplanan güvenirlik değeri .89 olarak bulunmuştur. Araştırmada, katılık ve cömertlikleri açısından puanlayıcılar arasında anlamlı fark olduğu tespit edilmiştir. Puanlayıcı yüzeyine ilişkin güvenirlik katsayısı .98 olarak hesaplanmıştır. Rasch modelindeki üçüncü yüzey bilimsel araştırma ödevlerinin değerlendirilmesinde kullanılan ölçütlerdir. Bu yüzeye ilişkin bulgular incelendiğinde, ölçütler arasında anlamlı fark olduğu tespit edilmiş ve ölçütlerin ödevlere ait farklı özellikleri ölçtüğü sonucuna ulaşılmıştır. Ölçüt yüzeyi için hesaplanan güvenirlik katsayısı ise .97 olarak elde edilmiştir.

Farrokhi ve Esfandiari (2011) tarafından yapılan araştırmada, yabancı dilde kompozisyon yazma performansının değerlendirilmesinde, puanlama işlemine karışan halo etkisi çok yüzeyli Rasch modeli ile incelenmiştir. Araştırmada,

öğrenciler tarafından yapılan öz değerlendirme ve akran değerlendirme ile öğretmenler tarafından yapılan değerlendirmeler halo etkisi açısından karşılaştırılmıştır. Puanlamalar, İngilizce öğretmenliği, İngiliz Edebiyatı ve Mütercim Tercümanlık bölümlerine kayıtlı toplamda 188 lisans öğrencisi ve altı İngilizce öğretmeni tarafından yapılmıştır. Yapılan puanlamalarda altılı derecelendirmeye sahip analitik bir rubrik kullanılmıştır. Öğretmenler tarafından yapılan değerlendirmeler, akran değerlendirme ve öz değerlendirme şeklinde toplamda 194 kişi tarafından yapılan puanlamalar halo etkisi açısından incelenmiştir. Araştırmadan elde edilen bulgular grup düzeyinde incelendiğinde, puanlama işlemine halo etkisinin karışmadığı belirlenmiştir. Ancak bulgular madde düzeyinde ele alındığında, her üç değerlendirme türü için de puanlama işlemine halo etkisinin karıştığı tespit edilmiş, 15 maddeden dördü için tüm puanlayıcı türlerinde halo etkisinin gözlendiği belirlenmiştir. Farklı puanlayıcı türleri birbirleri ile karşılaştırıldığında, öz değerlendirmenin akran değerlendirme ve öğretmenler tarafından yapılan değerlendirmeye göre halo etkisine daha açık olduğu belirlenmiştir.

Farrokhi, Esfandiari ve Vaez Dalili (2011) tarafından yapılan araştırmada, yabancı dilde kompozisyon yazma performansının değerlendirilmesinde, puanlama işlemine karışan merkeze yönelme etkisi çok yüzeyli Rasch modeli ile incelenmiştir. Araştırmada, öğretmenler tarafından yapılan değerlendirmeler ile öz değerlendirme ve akran değerlendirme merkeze yönelme etkisi açısından karşılaştırılmıştır. Araştırmada puanlamalar, İngilizce öğretmenliği, İngiliz Edebiyatı ve Mütercim Tercümanlık bölümlerine kayıtlı toplamda 188 lisans öğrencisi ve altı İngilizce öğretmeni tarafından yapılmıştır. Araştırmada öğretmenler tarafından yapılan değerlendirmeler ile akran değerlendirme ve öz değerlendirme şeklinde toplamda 194 kişi tarafından yapılan puanlamalar merkeze yönelme etkisi açısından incelenmiştir. Araştırma sonuçları incelendiğinde, her üç puanlayıcı türü için de grup ya da madde düzeyinde bir merkeze yönelme etkisinin gözlenmediği belirlenmiştir.

Semerci (2011a) tarafından yapılan araştırmada, doktora yeterlik sınavı çerçevesinde öğretim üyeleri tarafından yapılan değerlendirmeler ile akran ve öz değerlendirmelerin çok yüzeyli Rasch modeliyle analizinin yapılması amaçlanmıştır. Rasch ölçme modelinde üç yüzey yer almıştır. Bu yüzeyler, puanlayıcı olarak görev alan beş öğretim üyesi ve altı doktora öğrenci olmak üzere 11 jüri, 15 yeterlik alanı ve performansları değerlendirilen altı doktora öğrencisidir. Araştırmada yeterlik

düzeyleri açısından altı doktora öğrencisi arasında anlamlı fark bulunduğu tespit edilmiştir. Öğrenci yüzeyi için elde edilen güvenirlik katsayısı .91 olarak bulunmuştur. Puanlayıcı yüzeyine ilişkin bulgular incelendiğinde, puanlayıcılar arasında anlamlı farklılıkların bulunduğu saptanmıştır. Puanlayıcı yüzeyi için güvenirlik katsayısı .97 olarak hesaplanmıştır. Araştırmada ayrıca, öğrencilerin değerlendirilmesinde kullanılan yeterlilik alanları arasında anlamlı fark tespit edilmiştir. Yeterlik alanların güçlük düzeyleri arasında tespit edilen farkın .85 güvenirlikte elde edildiği belirlenmiştir.

Semerci (2011b) tarafından yapılan araştırmada, mikro öğretim uygulamaları çok yüzeyli Rasch modeliyle analiz edilmiştir. Araştırmanın çalışma grubunu, 2008-2009 Öğretim Yılı'nda Fırat Üniversitesi Bilgisayar ve Öğretim Teknolojileri Bölümü'nde özel öğretim yöntemleri dersini alan ve mikro öğretim uygulamasına katılan 32 öğretmen adayı oluşturmaktadır. Öğretmen adaylarının mikro öğretim uygulamaları dört puanlayıcı tarafından beşli derecelendirmeye sahip bir puanlama ölçeği kullanılarak değerlendirilmiştir. Elde edilen veriler çok yüzeyli Rasch modeline göre analiz edilmiştir. Analizlerde, öğretmen adayları, puanlayıcılar ve görevler olmak üzere üç yüzey işlem görmüştür. Araştırmada, öğretmen adaylarının mikro öğretim uygulamaları arasında anlamlı fark belirlenmiştir. Mikro öğretim uygulamalarına ilişkin öğretmen adayları arasında yapılan sıralamanın hangi güvenirlikte elde edildiğini gösteren güvenirlik katsayısı, .97 olarak hesaplanmıştır. Puanlayıcı yüzeyine ilişkin bulgular incelendiğinde, dört puanlayıcının katılık ve cömertlikleri yönüyle farklılık gösterdiği tespit edilmiştir. Puanlayıcı yüzeyine ilişkin güvenirlik katsayısı .99 olarak bulunmuştur. Görev yüzeyine ait bulgular incelendiğinde, mikro öğretim uygulamalarında kullanılan görevlerin güçlük düzeyleri arasında anlamlı farklılıkların olduğu belirlenmiştir. Araştırmada ayrıca, puanlayıcılardan bir kısmının yanlı puanlamalar yaptığı ve bazı öğretmen adaylarını puanlarken daha katı ya da daha cömert davrandığı belirlenmiştir.

Akın ve Baştürk (2012) tarafından yapılan araştırmada, güzel sanatlar ve spor lisesi öğrencilerinin keman eğitiminde kazanmış oldukları temel beceriler çok yüzeyli Rasch modeli ile değerlendirilmiştir. Üç farklı Güzel Sanatlar Lisesinde "Keman eğitimi" dersine devam eden 27 öğrenci çalışmaya dâhil edilmiştir. Araştırmada, keman eğitiminde uzman olan üç eğitimci beşli bir derecelendirmeye sahip bir rubrik kullanarak öğrencilerin performanslarını puanlamıştır. Araştırmadan elde edilen veriler, çok yüzeyli Rasch modeline göre analiz edilmiştir. Öğrenci

yüzeyine ilişkin bulgular, keman çalma becerileri açısından öğrenciler arasında anlamlı fark bulunduğu göstermiştir. Öğrenci yüzeyine ilişkin güvenirlik değeri .96 olarak hesaplanmıştır. Puanlayıcı yüzeyi için elde edilen bulgular incelendiğinde, katılık ve cömertlikleri açısından puanlayıcılar arasında anlamlı fark olmadığı belirlenmiştir. Puanlayıcı yüzeyi için hesaplanan güvenirlik katsayısı .41 olarak bulunmuştur. Keman çalma becerilerini değerlendirmede kullanılan ölçütlerden oluşan üçüncü yüzeye ilişkin bulgular incelendiğinde, güçlük düzeyleri açısından ölçütler arasında anlamlı fark olduğu belirlenmiştir. Bu yüzeye ilişkin güvenirlik katsayısı ise .97 olarak elde edilmiştir.

Hung, Chen ve Chen (2012) tarafından yapılan çalışmada, yaratıcılık çalışmaları kapsamında öğrencilerin tasarladıkları ürünlerin değerlendirilmesine karışan puanlayıcı etkileri incelenmiştir. Araştırmaya 113 öğrenci ve altı puanlayıcı dâhil edilmiştir. Puanlayıcılar dörtlü bir derecelendirmeye sahip sekiz boyuttan oluşan bir puanlama ölçeğini kullanarak değerlendirme yapmışlardır. Araştırmada, merkeze yönelme etkisi, halo etkisi, puanlayıcı katılığı ve cömertliği gibi olası yanlılık kaynaklarını tespit etmek için çok yüzeyli Rasch modeli kullanılmıştır. Elde edilen bulgulara göre, puanları önemli ölçüde etkileyen herhangi bir yanlılık kaynağının bulunmadığı belirlenmiştir. Performansları farklı olan öğrencilerin güvenilir bir biçimde ayırt edildiği sonucuna ulaşılmıştır.

Wiseman (2012a) tarafından yapılan araştırmada, öğrencilerin yabancı dil yazma performansını değerlendirmek amacıyla kullanılan analitik ve holistik rubriklerin karşılaştırılması amaçlanmıştır. Araştırmada 60 öğrencinin performansı, altılı derecelendirmeye sahip analitik ve holistik rubrik kullanılarak beş deneyimli puanlayıcı tarafından değerlendirilmiştir. Puanlayıcılar, holistik rubriğe göre değerlendirme yaparken, öğrencilerin performanslarına ilişkin genel izlenimden hareketle tek bir puan vermiştir. Analitik rubriğe göre yapılan değerlendirmelerde, öğrencilerin performansları; görevin yerine getirilmesi, konu belirleme, organizasyon, kelime ve dil kontrolü olmak üzere beş boyutta puanlanmıştır. Puanlayıcıların her bir boyut için verdikleri puanlar toplanarak öğrencilerin yazma yeteneklerine ilişkin genel bir puan elde edilmiştir. Puanlamalardan elde edilen veriler çok yüzeyli Rasch modeli ile analiz edilmiştir. Rasch analizi ile analitik ve holistik rubriklerin yazma yetenekleri farklı olan öğrencileri birbirinden ayırt etmede ne derece etkili olduğu incelenmiştir. Ayrıca rubriklerde kullanılan altılı derecelendirmenin etkililiği belirlenmeye çalışılmıştır. Araştırmada, puanlayıcıların

holistik rubrik ile puanlama yaparken genellikle puanlama ölçeğinin alt kategorilerini kullanma eğiliminde olduğu tespit edilmiştir. Bu noktadan hareketle, holistik rubrikte altılı bir derecelendirme yerine beşli bir derecelendirmenin kullanılmasının daha uygun olacağı ifade edilmiştir. Analitik rubriklerin, yazma yeteneği farklı olan öğrencileri birbirinden ayırt etmede oldukça etkili olduğu belirlenmiştir. Bu bulguya dayanarak, tanıma ve yerleştirmeye yönelik değerlendirmelerde analitik rubrik kullanılması daha doğru bir tercih olarak nitelendirilmiştir.

Wiseman (2012b) tarafından yapılan araştırmada, 39 öğrenci tarafından 39'u öykü ve 39'u inandırıcı yazı olmak üzere yabancı dilde yazılan toplam 78 kompozisyon sekiz puanlayıcı tarafından altılı derecelendirmeye sahip holistik ve analitik rubrik kullanılarak puanlanmıştır. Puanlayıcılar öğrencilerin performanslarını ilk olarak holistik rubrik ve daha sonra analitik rubrik kullanarak değerlendirmiştir. Rasch analizinden ve değerlendirmeler sırasında puanlayıcılar tarafından kaydedilen ses kayıtlarından elde edilen bulgulara göre; puanlayıcıların kültürel geçmişlerinin puanlama ile ilgili beklentilerini etkilediği ve yazma görevlerini puanlarken performans ölçütlerini uygulamadaki bireysel farklılıklara kaynaklık ettiği saptanmıştır. Araştırmadan elde edilen bulgular, puanlayıcının puanladığı kompozisyon ve/veya bu kompozisyonun yazarı ile olan uyumunun puanlamadaki davranışlarını etkilediğini ve puanlama katılığını azalttığını göstermiştir. Buna göre, puanlayıcıların iyi bildikleri ya da puanlama sırasında kendilerine yakın buldukları bireyleri değerlendirirken daha cömert davrandığı sonucuna ulaşılmıştır.

Tan (2013) tarafından yapılan araştırmada, yazma performansının değerlendirilmesinde kullanılan analitik rubriklerin geçerliği çok yüzeyli Rasch modeli ile incelenmiştir. Araştırmada, 10 öğrenci tarafından yazılan kompozisyonlar 13 puanlayıcı tarafından değerlendirilmiştir. Puanlayıcılar öğrencilerin performanslarını dört boyuttan oluşan analitik bir rubrik kullanarak değerlendirmiştir. Araştırmada çok yüzeyli Rasch modeli kullanılarak; puanlayıcı katılığı ve cömertliği, puanlayıcı tutarlığı, puanlayıcı yanlılığı, değerlendirmede kullanılan ölçütlerin güçlük düzeyi ve kullanılan puanlama ölçeğinin işlevselliği belirlenmeye çalışılmıştır. Elde edilen bulgulara göre, puanlayıcıların analitik rubrikleri başarılı bir biçimde kullandığı ve yazma yeteneği farklı olan öğrencileri tutarlı bir biçimde ayırt edebildiği belirlenmiştir. Araştırmada ayrıca, puanlayıcıların puanlama konusundaki deneyimlerinin yaptıkları değerlendirmeler üzerinde etkili olmadığı sonucuna ulaşılmıştır.

Qinghui (2013) tarafından yapılan çalışmada, öğretmenlerin öğretim uygulamalarının çok yüzeyli Rasch modeli ile değerlendirilmesi amaçlanmıştır. Çalışmaya beş öğretmen ve 43 üniversite öğrencisi dâhil edilmiştir. Öğretmenlerin Çinceyi öğretmede ne kadar yetkin oldukları 43 öğrenci tarafından değerlendirilmiştir. Araştırmadan elde edilen bulgulara göre; Çinceyi öğretme yetkinlikleri farklı olan öğretmenlerin birbirinden ayırt edilmesinde puanlama ölçeğinin etkin bir biçimde çalıştığı, öğretmenlerin Çince öğretme yetkinliklerinde anlamlı farklılıklar olduğu, puanlayıcıların puanlama ölçeğini tutarlı bir biçimde kullandığı belirlenmiştir. Araştırmada öğretmen (puanlanan kişiler) yüzeyi için hesaplanan güvenirlik katsayısının .99 olduğu ve öğretim performansları farklı olan öğretmenlerin güvenilir bir biçimde ayırt edildiği saptanmıştır. Öğretmenlerden biri için öğrencilerin yaptıkları puanlamalarda tutarsızlıklar olduğu ve bu öğretmen için hesaplanan dış uyum kareler ortalamasının kabul edilebilir aralığının dışında yer aldığı belirlenmiştir. Puanlayıcı yüzeyine (öğretmenleri puanlayan öğrenciler) ilişkin bulgular incelendiğinde, puanlayıcıların katılık ve cömertliklerinde anlamlı farklılıklar olduğu saptanmıştır. Bu yüzey için hesaplanan ayırma indeksi güvenirliği .89 olarak hesaplanmıştır. Puanlayıcılar arası hesaplanan mutlak uyum oranı %53 olarak bulunmuş ve bu değerin %52.6'lık beklenen uyum değerinden yüksek olduğu tespit edilmiştir. Puanlayıcılardan ikisi için dış uyum kareler ortalamasının kabul edilebilir aralığın üstünde yer aldığı ve bu puanlayıcıların tutarsız puanlamalar yaptığı belirlenmiştir. Başka iki puanlayıcı için ise dış uyum kareler ortalamasının kabul edilebilir aralığın altında olduğu tespit edilmiştir. Bu puanlayıcıların öğretim performansları farklı olan öğretmenleri birbirinden ayırt edemediği ve tüm öğretmenlere benzer puanlar verdiği saptanmıştır. Öğrencilerin, öğretmenlerin öğretim performansını değerlendirirken yanlı puanlamalar yapıp yapmadığını belirlemek için yanlılık analizine ilişkin bulgular incelenmiştir. Araştırmada puanlama yapan 43 öğrenci ve puanlanan beş öğretmen bulunmaktadır. Buna göre 43×5=215 puanlama yapılmıştır. Bu 215 puanlamadan 10'u için öğretmen ve öğrenciler arasındaki etkileşime ilişkin *t* değerinin istatistiksel olarak anlamlı olduğu belirlenmiştir. Yanlılık tespit edilen puanlama sayısının toplam puanlama sayısına oranı %4.7 olarak hesaplanmıştır. Bu oran toplam puanlama sayısının %5'inden az olduğundan yapılan puanlamaların kabul edilebilir aralıkta yer aldığı sonucuna ulaşılmıştır.

Güler (2014) tarafından yapılan araştırmada açık uçlu istatistik sorularına verilen yanıtların çok yüzeyli Rasch modeline göre incelenmesi amaçlanmıştır. Araştırmada, 55 öğrencinin 10 açık uçlu istatistik sorusuna verdiği yanıtlar ortak bir puanlama anahtarı kullanarak değerlendirme yapan üç puanlayıcı tarafından puanlanmıştır. Dolayısıyla araştırmaya birey, madde ve puanlayıcı olmak üzere üç yüzey dâhil edilmiştir. Araştırmada birey yüzeyi için hesaplanan güvenirlik katsayısı .79 ve madde yüzeyi için hesaplanan güvenirlik katsayısı .90 olarak bulunmuştur. Puanlayıcı yüzeyi için hesaplanan güvenirlik katsayısı ise .00 olarak bulunmuştur. Puanlamadaki katılık/cömertlikleri açısından puanlayıcılar arasında anlamlı fark bulunmadığı tespit edilmiştir.

Puanlayıcı etkilerinin çok yüzeyli Rasch modeli ile incelendiği çalışmalara bakıldığında, çok yüzeyli Rasch modelinin klasik test kuramının ötesinde birçok avantaj sunduğu görülmektedir. Bu avantajlar, puanlayıcı etkilerinin belirlenmesine yönelik çalışmalarda çok yüzeyli Rasch modelini öncelikle olarak başvurulması gereken bir ölçme kuramı haline getirmektedir.

ÜÇÜNCÜ BÖLÜM

MATERYAL VE YÖNTEM

Bu bölümde; çalışmanın türü, katılımcılar, çalışmada kullanılan veri toplama araçları, araştırma süresince takip edilen işlemler ve verilerin analizinde kullanılan istatistiksel teknikler açıklanmıştır.

3.1. ÇALIŞMANIN TÜRÜ

Bu çalışmada, standart ve SOLO taksonomisine dayalı rubrikler ile puanlanan açık uçlu matematik sorularında puanlayıcı etkilerinin incelenmesi amaçlanmaktadır. Çalışma hem temel hem de uygulamalı bir araştırma niteliği taşımaktadır. Temel araştırmalar kuram geliştirme ve bir kuramın test edilmesine yönelik çalışmalardır (Kaptan, 1998). SOLO taksonomisine dayalı rubriklerin puanlayıcı hatalarını kontrol altına alma konusundaki etkililiğinin test edilmesi yönüyle bu çalışmanın temel bir araştırma niteliği taşıdığı düşünülmektedir.

Uygulamalı araştırmalar ise, kuram ile uygulama arasındaki boşluğun doldurulmasına hizmet eden ve güncel problemlere çözüm yolları bulma amacı taşıyan çalışmalardır (Kaptan, 1998). Açık uçlu soruların puanlayıcı kaynaklı faktörlerden etkilenmeye açık doğası gerek sınıf içi değerlendirmeler gerekse de geniş ölçekli sınavlar için oldukça önemli bir problemdir. Açık uçlu soruların puanlanması ile ilgili problemlere çözüm yolları bulmayı amaç edinmesi açısından çalışmanın uygulamalı bir araştırma niteliği taşıdığı söylenebilir.

3.2. KATILIMCILAR

Araştırmanın katılımcıları, açık uçlu matematik sorularının uygulandığı öğrenciler ile öğrenciler tarafından cevaplanan sınav kâğıtlarını değerlendiren puanlayıcılardan oluşmaktadır.

3.2.1. Öğrenci Grubu

Araştırma, 2013-2014 Eğitim-Öğretim Yılı Bahar Dönemi'nde Diyarbakır ili merkez ilçelerdeki dört farklı ilköğretim okulunun sekizinci sınıfına devam eden 104 öğrenci üzerinde yürütülmüştür. Öğrencilerin 46'sı (%44.20) kız ve 58'i (%55.80) erkektir.

3.2.2. Puanlayıcılar

Araştırmada, öğrenciler tarafından cevaplanan açık uçlu matematik sorularının değerlendirilmesinde yedi puanlayıcı görev almıştır. Puanlayıcılardan üçü kadın ve dördü erkektir. Puanlayıcılar araştırmaya gönüllü olarak katılmışlardır. Puanlayıcıların demografik özelliklerine ilişkin bilgiler Tablo 3.1'de sunulmuştur.

Tablo 3.1. Puanlayıcılara ilişkin demografik bilgiler

Puanlayıcı	Cinsiyet	Yaş	Öğretmenlikteki Görev Süresi	Eğitim Durumu
P1	Kadın	22	-	İlköğretim matematik öğretmenliği mezunudur ve matematik eğitimi alanında yüksek lisans yapmaktadır.
P2	Kadın	22	7 ay	İlköğretim matematik öğretmenliği mezunudur ve eğitimde ölçme değerlendirme alanında yüksek lisans yapmaktadır.
P3	Kadın	23	7 ay	İlköğretim matematik öğretmenliği mezunudur ve eğitimde ölçme değerlendirme alanında yüksek lisans yapmaktadır.
P4	Erkek	26	2 yıl	İlköğretim matematik öğretmenliği mezunudur ve eğitimde ölçme değerlendirme alanında yüksek lisans yapmaktadır.
P5	Erkek	25	2 yıl	İlköğretim matematik öğretmenliği mezunudur ve eğitimde ölçme değerlendirme alanında yüksek lisans yapmaktadır.
P6	Erkek	25	7 ay	İlköğretim matematik öğretmenliği mezunudur ve eğitimde ölçme değerlendirme alanında yüksek lisans yapmaktadır.
P7	Erkek	26	3 yıl	Bilgisayar bilimleri ağırlıklı matematik alanından mezun olup formasyon eğitimi almıştır. Eğitimde ölçme değerlendirme alanında yüksek lisans yapmaktadır.

3.3. VERİ TOPLAMA ARAÇLARI

Araştırma verilerinin toplanmasında; açık uçlu sorulardan oluşan matematik başarı testi, standart rubrikler, SOLO taksonomisine dayalı rubrikler ile puanlayıcıların standart ve SOLO taksonomisine dayalı rubrikler hakkındaki düşüncelerini belirlemek amacıyla hazırlanmış anketler kullanılmıştır.

3.3.1. Açık Uçlu Sorulardan Oluşan Matematik Başarı Testi

Açık uçlu sorulardan oluşan matematik başarı testi araştırmacı tarafından geliştirilmiştir. Testin geliştirilmesi çeşitli aşamalarda gerçekleşmiştir. Bu aşamalar aşağıda açıklanmıştır.

Testte yer alacak maddelerin yazılması: Maddeler yazılmadan önce, Matematik Dersi (6-8. Sınıflar) Öğretim Programı'nda yer alan öğrenme alanları ve kazanımlar incelenmiştir. Daha sonra; sayılar, geometri, cebir, ölçme, olasılık ve istatistik öğrenme alanlarına yönelik toplam 18 açık uçlu soru hazırlanmıştır.

Hazırlanan maddeler için uzman görüşü alınması: Madde yazımının ardından hazırlanan soruların sekizinci sınıf düzeyine uygunluğunu ve anlaşılırlığını değerlendirmek üzere 10 uzmandan görüş alınmıştır. Görüşleri alınan uzmanların demografik özelliklerine ilişkin bilgiler Tablo 3.2'de sunulmuştur.

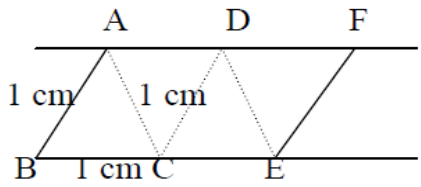
Tablo 3.2. Hazırlanan açık uçlu soruları anlaşılırlık ve sekizinci sınıf düzeyine uygunluk açısından değerlendiren uzmanlara ilişkin demografik özellikler

Uzmanlar	Cinsiyet	Eğitim Durumu
1	Erkek	Matematik eğitimi alanında doçenttir.
2	Erkek	Matematik eğitimi alanında doçenttir.
3	Kadın	İlköğretim matematik öğretmenliği alanından mezundur. Matematik eğitimi alanında yüksek lisans eğitimini tamamlamıştır. Eğitim programları ve öğretim alanında doktora eğitimine devam etmektedir.
4	Erkek	İlköğretim matematik öğretmenliği alanından mezundur. Eğitim programları ve öğretim alanında yüksek lisansını tamamlamış ve aynı alanda doktora eğitimine devam etmektedir.
5	Kadın	İlköğretim matematik öğretmenliği alanından mezundur. Eğitim programları ve öğretim alanında yüksek lisans eğitimini tamamlamıştır, aynı alanda doktora eğitimine devam etmektedir.
6	Erkek	İlköğretim matematik öğretmenliği alanından mezundur. Eğitimde ölçme değerlendirme alanında yüksek lisansını tamamlamıştır.
7	Erkek	İlköğretim matematik öğretmenliği alanından mezundur. Matematik eğitimi alanında yüksek lisans eğitimine devam etmektedir.
8	Erkek	İlköğretim matematik öğretmenliği alanından mezundur. Eğitimde ölçme değerlendirme alanında yüksek lisans eğitimine devam etmektedir.
9	Kadın	İlköğretim matematik öğretmenliği alanından mezundur. Eğitimde ölçme değerlendirme alanında yüksek lisans eğitimine devam etmektedir.
10	Kadın	İlköğretim matematik öğretmenliği alanından mezundur. Eğitimde ölçme değerlendirme alanında yüksek lisans eğitimine devam etmektedir.

Uzmanlar; *Madde, bu haliyle ölçme aracında yer alabilir* (3), *Madde düzeltildikten sonra ölçme aracında yer alabilir* (2) ve *Maddenin ölçme aracından çıkarılması gerekir* (1) şeklinde üçlü derecelendirmeye sahip bir ölçek kullanarak

hazırlanan maddeleri değerlendirmiştir. Araştırmada, öğrencilerin herhangi bir matematik konusundaki başarılarının belirlenmesiyle ilgilenilmemektedir. Dolayısıyla uzmanlar kapsam geçerliği açısından bir değerlendirme yapmamıştır.

Uzman görüşleri doğrultusunda maddelerin gözden geçirilmesi: Uzmanlardan alınan görüşler doğrultusunda anlaşılırlık açısından problem oluşturabileceği ifade edilen ya da sekizinci sınıf düzeyine uygun olmadığı düşünülen altı madde ölçme aracından çıkarılmıştır. Örneğin, ölçme öğrenme alanına ait "Çokgenlerin kenar uzunlukları ile çevre uzunluğu arasındaki ilişkiyi açıklar." kazanımını değerlendirmeye yönelik aşağıdaki madde, anlaşılırlık açısından problem oluşturabileceği gerekçesiyle ölçme aracından çıkarılmıştır.



Yan tarafta, kenar uzunluğu 1 cm olan eşkenar üçgenlerden oluşan bir tren gösterilmiştir. Trenin çevresinin uzunluğu, treni oluşturan üçgen sayısına göre değişmektedir. Treni oluşturan üçgen sayısı ile trenin çevre uzunluğu arasındaki ilişkiyi cebirsel olarak ifade ediniz.

Yukarıdaki madde için uzmanlardan ikisi "*Trenin üç boyutlu bir şekil olduğunu ve trenin çevresi ile ne kastedildiğinin açık olmadığını*" belirtmiştir. Bu nedenle, madde ölçme aracından çıkarılmıştır. Uzman görüşleri doğrultusunda bazı maddeler ölçme aracından çıkarılırken; bazı maddelerin ifade ediliş şekillerinde değişikliğe gidilmiştir. Örneğin; uzmanlardan üçü "a+b=8 ve b<a olduğuna göre, b'nin alabileceği değerler için ne söyleyebilirsiniz?" şeklinde ifade edilen soruda a ve b'nin ait olduğu sayı kümelerinin belirtilmesi gerektiğini ifade etmiştir. Dolayısıyla, bu madde "a ve b birer reel sayı, a+b=8 ve b<a olduğuna göre, b'nin alabileceği değerler için ne söyleyebilirsiniz? şeklinde yeniden düzenlenmiştir. Uzman görüşleri doğrultusunda değişiklik yapılan bir diğer madde cebir öğrenme alanında bulunan "Belirli durumlara uygun cebirsel ifade yazar" kazanımını değerlendirmeye yöneliktir. Bu kazanımı değerlendirmek amacıyla yazılan madde; "A bir matematiksel denklem makinesi olmak üzere, A makinesine bir sayı eklendiğinde" ifadesi ile başlamaktadır. 10 uzmandan yedisi bu madde de *eklendiğinde* sözcüğü yerine *konulduğunda* sözcüğünün kullanılmasını daha doğru bir tercih olarak belirtmiştir. Bunun üzerine bu madde, "A bir matematiksel denklem makinesi olmak üzere A makinesine bir sayı konulduğunda" şeklinde yeniden ifade edilmiştir.

Uzman görüşleri doğrultusunda değişiklik yapılan başka bir madde yine cebir öğrenme alanına yöneliktir. Bu madde; "Ahmet ve Elif oynadıkları oyunda ellerindeki eşit boydaki kürdanları kullanarak beşgen şeklinde ve yan yana olacak biçimde evler yapmaktadır. Ahmet ve Elif, yaptıkları ev sayısı ile ellerindeki kürdan sayısı arasında nasıl bir ilişki olacağını merak etmektedir" ifadelerini içermektedir. Madde için uzmanlardan ikisi böyle bir oyun oynayan çocukların bu türden bir ilişkiyi merak etmeyeceğini belirtmiş ve sorunun doğrudan öğrenciye yöneltilmesinin daha uygun olacağını dile getirmiştir. Bu noktadan hareketle söz konusu madde, "Ahmet ve Elif'in ellerindeki kürdan sayısı ile yaptıkları ev sayısı arasında nasıl bir ilişki vardır" şeklinde yeniden düzenlenmiş ve doğrudan soruyu çözecek öğrenciye yöneltilmiştir.

Küçük bir öğrenci grubu üzerinde ön uygulama yapılması: Uzman görüşlerinden sonra; altı maddenin ölçekten çıkarılması ve getirilen öneriler doğrultusunda beş maddede gerekli değişikliklerin yapılmasının ardından, 12 maddeden oluşan bir test elde edilmiştir. Teste son şeklini vermeden önce küçük bir öğrenci grubu üzerinde ön uygulama yapılmıştır. Bu amaçla, hazırlanan test sekizinci sınıfa devam eden yedisi kız ve altısı erkek olmak üzere toplam 13 öğrenciye uygulanmıştır. Ön uygulama ile testte yer alan maddeler ve testin başında sunulan yönerge hakkında öğrenci görüşlerinin belirlenmesi amaçlanmıştır. Öğrencilerden, anlamakta güçlük çektikleri ifadeler olması durumunda bunları her bir sorunun altında verilen "*Anlayamadığım İfadeler*" başlığı altına not etmeleri istenmiştir. Ön uygulamanın yapıldığı öğrenci grubunun düşük, orta ve yüksek başarı düzeylerini temsil edebilecek nitelikte olmasına özellikle dikkat edilmiştir. Ön uygulamanın ardından, anlaşılırlığında herhangi bir problem olmadığı görülen 10 madde belirlenmiştir. Ancak esas uygulamada 10 maddelik bir testin süre açısından sıkıntı yaratabileceği düşünülerek testteki soru sayısı sekize düşürülmüştür. Daha sonra, sekiz maddeden oluşan başarı testinin uygulama süresi hakkında geri bildirim almak ve maddeleri anlaşılırlık açısından bir kez daha gözden geçirmek için sekizinci sınıfa devam eden 15 öğrenci (yedi kız ve sekiz erkek) üzerinde ikinci bir ön uygulama yapılmıştır. Ön uygulamada öğrencilerin testin başında sunulan yönergede ya da test maddelerinde anlamakta güçlük çektikleri herhangi bir ifadeye rastlanmamıştır. Ön uygulamanın yapıldığı grupta testi en kısa sürede tamamlayan öğrenci ile en geç tamamlayan öğrencinin kullandıkları süreler dikkate alınarak testin uygulama süresi

bir ders saati (40 dakika) olarak belirlenmiştir. Ön uygulamadan sonra nihai şekli verilen başarı testindeki soruların altı tanesi SOLO taksonomisinin ilişkisel yapı düzeyine yönelik maddelerdir. Kalan iki soru ise soyutlanmış yapı düzeyine karşılık gelmektedir. Dolayısıyla, testteki sekiz sorudan altısı için verilen cevaplar; yapı öncesi düzey ile ilişkisel yapı düzeyi aralığında değişebilirken; diğer iki soru için verilen cevaplar yapı öncesi düzey ile soyutlanmış yapı düzeyi aralığında uzanabilmektedir. Açık uçlu sorulardan oluşan matematik başarı testinin nihai formu Ek-1'de sunulmuştur.

3.3.2. Rubrikler

Öğrencilerin açık uçlu matematik sorularına verdikleri cevapların puanlanmasında standart ve SOLO taksonomisine dayalı rubriklerden yararlanılmıştır. Hem standart hem de SOLO taksonomisine dayalı rubrikler göreve özel holistik rubrik şeklinde geliştirilmiştir. Diğer bir ifadeyle, matematik başarı testinde yer alan her bir sorunun puanlanmasında ayrı bir standart ve SOLO taksonomisine dayalı rubrik kullanılmıştır. Standart ve SOLO taksonomisine dayalı rubriklerin puanlayıcı etkileri açısından karşılaştırılmasında, rubriklerde kullanılan derecelendirmelerin farklı olmasından kaynaklı herhangi bir etki oluşmaması için iki rubrik türünde eş bir derecelendirme benimsenmiştir. Matematik başarı testinde yer alan ilk altı soru için hem standart hem de SOLO taksonomisine dayalı rubriklerde dörtlü bir derecelendirme esas alınmıştır. Testin yedi ve sekiz numaralı soruları için ise her iki rubrik türünde de beşli bir derecelendirme kullanılmıştır.

3.3.2.1. Standart Rubrikler

Açık uçlu sorulardan oluşan matematik testindeki maddelerin her biri için bir rubrik olmak üzere sekiz standart rubrik geliştirilmiştir. Geliştirilen rubriklerin altı tanesinde dörtlü bir derecelendirme kullanılmıştır. Alt amaçlardan oluşan iki soru için ise beşli derecelendirme esas alınmıştır. Daha sonra geliştirilen rubrikler beş uzmanın görüşüne sunulmuştur. Uzmanlar rubrikleri; anlaşılırlık, ölçülmek istenen yapıyı tam olarak yansıtıp yansıtmaması ve puanlama kategorilerinin uygunluğu açısından değerlendirmiştir. Uzmanlar rubrikleri değerlendirirken üçlü derecelendirmeye sahip bir form kullanmıştır (Ek-4). Standart rubrikler hakkında görüşüne başvurulan uzmanların demografik özelliklerine ilişkin bilgiler Tablo 3.3'te gösterilmiştir.

Tablo 3.3. Standart rubrikler hakkında görüşüne başvurulan uzmanların demografik özelliklerine ilişkin bilgiler

Uzman	Cinsiyet	Eğitim Durumu
1	Erkek	Eğitimde ölçme ve değerlendirme alanında doçenttir.
2	Erkek	Sınıf öğretmenliği alanında doçenttir.
3	Kadın	Eğitim programları ve öğretim alanında yüksek lisans ve doktora eğitimini tamamlamıştır.
4	Erkek	İlköğretim matematik öğretmenliği alanından mezun olmuştur. Eğitim programları ve öğretim alanında yüksek lisansını tamamlamış ve aynı alanda doktora eğitimine devam etmektedir.
5	Erkek	Ortaöğretim matematik öğretmenliği alanından mezun olmuştur. Matematik eğitimi alanında yüksek lisansını tamamlamış ve aynı alanda doktora eğitimine devam etmektedir.

Uzmanlardan alınan görüşler; *i)* puanlama kategorilerinin iyi tanımlandığını, *ii)* puanlama kategorileri arasındaki farkların açık olduğunu, *iii)* rubriklerin her nitelikteki öğrenci grubunu ölçmek için kullanılabileceğini, i*v)* puanlama ölçütlerinin soru ile ölçülmek istenen özelliğin bütün yönlerini yansıttığını, *v)* puanlama ölçütlerinin soru ile ölçülmek istenen özellik dışında herhangi bir değerlendirme ölçütü içermediğini ve *vi)* rubriklerdeki ifadelerin anlaşılır olduğunu göstermiştir. Dolayısıyla uzman görüşleri, hazırlanan standart rubriklerin kullanıma hazır olduğunu ve herhangi bir değişikliğe ihtiyaç duyulmadığını ortaya koymuştur. Açık uçlu matematik sorularının puanlanmasında kullanılan standart rubrikler Ek-2'de verilmiştir.

3.3.2.2. SOLO Taksonomisine Dayalı Rubrikler

Açık uçlu sorulardan oluşan matematik testindeki maddelerin her biri için bir rubrik olmak üzere SOLO taksonomisine dayalı sekiz rubrik geliştirilmiştir. Matematik testindeki soruların altısı ilişkisel yapı düzeyine yöneliktir. Başka bir ifadeyle, matematik başarı testindeki sekiz maddenin altısı öğrencilerin en fazla ilişkisel yapı düzeyinde yanıt verebileceği sorulardır. Dolayısıyla bu sorulara yönelik olarak geliştirilen rubriklerde *yapı öncesi* (0), *tek yönlü yapı* (1), *çok yönlü yapı* (2) ve *ilişkisel yapı* (3) şeklinde dörtlü bir derecelendirme kullanılmıştır. Testteki diğer iki soru ise soyutlanmış yapı düzeyine yöneliktir. Öğrencilerin bu sorulara verecekleri cevaplar yapı öncesi düzeyden soyutlanmış yapı düzeyine kadar uzanabilmektedir. Buna bağlı olarak, bu iki sorunun puanlanmasında kullanılmak üzere geliştirilen SOLO taksonomisine dayalı rubriklerde *yapı öncesi* (0), *tek yönlü yapı* (1), *çok yönlü yapı* (2), *ilişkisel yapı* (3) ve *soyutlanmış yapı* (4) olmak üzere

beşli bir derecelendirme esas alınmıştır. SOLO taksonomine dayalı rubrikler geliştirildikten sonra, dört uzmanın görüşüne sunulmuştur. SOLO taksonomisine dayalı rubrikler hakkında görüşüne başvurulan uzmanların demografik özelliklerine ilişkin bilgiler Tablo 3.4'te verilmiştir.

Tablo 3.4. SOLO taksonomisine dayalı rubrikler hakkında görüşüne başvurulan uzmanların demografik özelliklerine ilişkin bilgiler

Uzman	Cinsiyet	Eğitim Durumu
1	Erkek	Eğitimde ölçme ve değerlendirme alanında doçenttir.
2	Kadın	Eğitim programları ve öğretim alanında yüksek lisans ve doktora eğitimini tamamlamıştır.
3	Erkek	Matematik öğretmenliği alanından mezun olmuştur. Matematik eğitimi alanında yüksek lisans eğitimini tamamlamıştır.
4	Erkek	İlköğretim matematik öğretmenliği alanından mezundur. Matematik eğitimi alanında yüksek lisans eğitimine devam etmektedir ve matematik öğretmeni olarak görev yapmaktadır.

Uzmanlardan alınan görüşler; *i)* puanlama kategorilerinin SOLO taksonomisinin düzeylerine uygun olarak tanımlandığını, *ii)* puanlama kategorileri arasındaki farkların açık olduğunu, *iii)* rubriklerin her nitelikteki öğrenci grubunu ölçmek için kullanılabileceğini, i*v)* puanlama ölçütlerinin soru ile ölçülmek istenen özelliğin bütün yönlerini yansıttığını, *v)* puanlama ölçütlerinin soru ile ölçülmek istenen özellik dışında herhangi bir değerlendirme ölçütü içermediğini ve *vi)* rubriklerdeki ifadelerin anlaşılır olduğunu göstermiştir. Bu nedenle uzman görüşleri; SOLO taksonomisine dayalı rubriklerin kullanıma hazır olduğu ve herhangi bir değişikliğe ihtiyaç duyulmadığı şeklinde yorumlanmıştır. Açık uçlu matematik sorularının puanlanmasında kullanılan SOLO taksonomisine dayalı rubrikler Ek-3'te verilmiştir.

3.3.3. Puanlayıcıların Standart ve SOLO Taksonomisine Dayalı Rubrikler Hakkındaki Düşüncelerini Belirlemede Kullanılan Anketler

Puanlayıcıların standart rubrikler hakkındaki görüşlerini belirlemek için Standart Rubrikler ile İlgili Düşünceler Anketinden (SRDA), SOLO taksonomisine dayalı rubrikler hakkındaki görüşlerini belirlemek için ise SOLO Taksonomisine Dayalı Rubrikler ile İlgili Düşünceler Anketinden (STDRDA) yararlanılmıştır (Ek-5 ve Ek-6). SRDA ve STDRDA'nın hazırlanması sürecinde takip edilen işlemler aşağıda sunulmuştur.

3.3.3.1. Standart Rubrik ile İlgili Düşünceler Anketi (SRDA)

Puanlayıcıların standart rubrikler hakkındaki görüşlerini belirmek için iki bölümden oluşan bir anket hazırlanmıştır. Anketin birinci bölümünde kapalı uçlu (yapılandırılmış) beş madde yer almaktadır. İlgili literatür taranarak oluşturulan söz konusu beş madde için beşli bir derecelendirme kullanılmıştır. Daha sonra maddelerin kapsamı ve anlaşılırlığı hakkında iki ölçme değerlendirme uzmanından görüş alınmıştır. Uzman görüşleri; ilk maddenin daha açık bir biçimde ifade edilmesi gerektiğini, diğer dört maddede ise herhangi bir değişikliğe ihtiyaç duyulmadığını göstermiştir. Uzmanlar, "Standart Rubriklerin Objektifliği [*Çok Düşük* (1)→ *Çok Yüksek* (5)]" şeklinde ifade edilen birinci maddede, puanlayıcılar arasında objektiflik kavramına ilişkin ortak bir algı oluşmasını sağlamak için objektiflik ile ne kastedildiğinin açıklanmasını önermiştir. Bu öneri doğrultusunda, "Standart Rubrikleri Objektifliği" ifadesinin yanına "Puanlama işlemi kim tarafından yapılırsa yapılsın aynı sonuca ulaşılması" şeklinde bir açıklama eklenmiştir. Uzman görüşlerinden hareketle, gerekli değişikliklerin yapılmasının ardından SRDA'nın kapalı uçlu sorulardan oluşan birinci bölümü kullanıma hazır hale gelmiştir.

Anketin ikinci bölümü puanlayıcıların standart rubrikler hakkındaki görüşlerini belirlemeye yönelik açık uçlu (yapılandırılmamış) maddelerden oluşmaktadır. Açık uçlu maddeler hazırlanırken, kolay anlaşılabilir nitelikte sorular sormaya özen gösterilmiş; yoruma açık, çok boyutlu ve yönlendirici sorulardan (Büyüköztürk, Çakmak, Akgün, Karadeniz ve Demirel, 2010; Yıldırım ve Şimşek, 2011) kaçınılmıştır. Ayrıca evet/hayır gibi tek bir kelime ile cevaplanabilecek sorulardan uzak durulmuştur. Puanlayıcıların görüşlerini açık bir biçimde ifade etmesini sağlayacak sorulara yer verilmiştir. Örneğin; puanlayıcıların standart rubriklerin objektifliği hakkındaki düşüncelerini belirlemek için "Standart rubriklerin objektif olduğunu düşünüyor musunuz" şeklinde bir madde yerine "Standart rubriklerin objektifliği hakkında ne düşünüyorsunuz" biçiminde bir maddeye yer verilmiştir. Anketin ikinci bölümü için; standart rubriklerin objektifliği, kullanım kolaylığı, öğrencilere güçlü ve zayıf olduğu noktalar hakkında geri bildirim vermedeki etkililiği hakkında toplam altı açık uçlu madde hazırlanmıştır. Daha sonra hazırlanan açık uçlu maddeler için eğitim programları ve öğretim, ölçme-değerlendirme ve sınıf öğretmenliği alanlarından birer uzman olmak üzere üç uzmandan görüş alınmıştır. Uzman görüşleri doğrultusunda açık uçlu maddeler

gözden geçirilerek, gerekli değişiklikler yapılmıştır. Böylece anketin açık uçlu maddelerden oluşan ikinci bölümü de uygulamaya hazır hale gelmiştir.

3.3.3.2. SOLO Taksonomisine Dayalı Rubrikler ile İlgili Düşünceler Anketi (STDRDA)

Puanlayıcıların SOLO taksonomisine dayalı rubrikler hakkındaki görüşlerini belirlemeye yönelik sorular, SRDA'da bulunan kapalı ve açık uçlu maddelerdeki "standart rubrik" ifadeleri "SOLO taksonomisine dayalı rubrik" şeklinde değiştirilerek elde edilmiştir. Örneğin; anketin kapalı uçlu maddelerden oluşan birinci bölümünde yer alan "Standart rubriklerin öğrenciye güçlü ve zayıf olduğu noktalar hakkında geri bildirimler sunabilme özelliği [*Çok Düşük* (1)→ *Çok Yüksek* (5)]" sorusu "SOLO taksonomisine dayalı rubriklerin öğrenciye güçlü ve zayıf olduğu noktalar hakkında geri bildirimler sunabilme özelliği [*Çok Düşük* (1)→ *Çok Yüksek* (5)]" şeklinde değiştirilmiştir. Benzer değişiklikler anketin açık uçlu maddelerden oluşan ikinci bölümünde de yapılmıştır. Söz gelimi; "Standart rubrikler kullanılmadan önce bu rubriklere yönelik bir eğitim verilmesinin gerekliliği hakkında ne düşünüyorsunuz?" sorusu, "SOLO taksonomisine dayalı rubrikler kullanılmadan önce bu rubriklere yönelik bir eğitim verilmesinin gerekliliği hakkında ne düşünüyorsunuz?" biçiminde değiştirilmiştir. SRDA'dan farklı olarak, STDRDA'ya "Açık uçlu matematik soruların puanlanmasında standart ve SOLO taksonomisine dayalı rubriklerden hangisini tercih edersiniz. Neden?" şeklinde bir madde eklenmiştir. Dolayısıyla SRDA'nın ikinci bölümünde altı açık uçlu madde yer alırken; STDRDA'nın ikinci bölümünde yedi açık uçlu madde bulunmaktadır. Puanlayıcıların SOLO taksonomisi ile ilgili görüşlerini belirlemeye yönelik sorular, standart rubrikler ile ilgili sorular temele alınarak hazırlandığından bu maddeler için ayrıca uzman görüşüne başvurulmamıştır. Yalnızca, SRDA'da yer almayan ancak STDRDA'da bulunan açık uçlu yedinci madde için uzman görüşüne ihtiyaç duyulmuştur. Bu madde, standart rubrikler için hazırlanan açık uçlu maddeleri inceleyen üç uzmana sunulmuştur. Uzman görüşleri, maddenin anlaşılır olduğunu ve yönlendirici herhangi bir ifade içermediğini ortaya koymuştur. Böylelikle, STDRDA'da uygulamaya hazır hale gelmiştir.

3.4. İŞLEM

Araştırmanın verileri, 2013-2014 Öğretim Yılı Bahar Dönemi'nde toplanmıştır. Araştırmacı tarafından geliştirilen matematik başarı testi, sınıf ortamında öğrencilere uygulanmıştır. Uygulama sırasında, araştırmacının sınıfta bulunmasına özen gösterilmiştir. Veri toplama aracı uygulanmadan önce öğrenciler araştırmanın amacı hakkında bilgilendirilmiştir. Öğrencilere toplanan verilerin yalnızca araştırmanın amacı için kullanılacağı, başka herhangi bir kurum ya da kişi ile paylaşılmayacağı ifade edilmiştir. Yine uygulamadan önce araştırmaya katılmanın zorunlu olmadığı ifade edilerek araştırma grubunun yalnızca gönüllü öğrencilerden oluşması sağlanmıştır. Ayrıca, matematik başarı testindeki soruları içtenlikle yanıtlamalarının geçerli ve güvenilir sonuçlar elde edilebilmesi için son derece önemli olduğu araştırmacı tarafından öğrencilere hatırlatılmıştır. Öğrencilerden soruların çözümüne yönelik yaptıkları işlemleri açık bir biçimde yazmaları rica edilmiş ve test için kendilerine 40 dakika süre verileceği belirtilmiştir. Başarı testinin ilk bölümünde; cinsiyet, yaş, birinci dönem matematik dersi karne notu ve birinci dönem yapılan matematik dersi ortak sınavındaki doğru sayısı değişkenlerine yer verilerek öğrencilerin demografik özelliklerine ilişkin bilgiler toplanmıştır. Öğrencilerin büyük çoğunluğu yaklaşık 30 ile 35 dakika içinde başarı testini tamamlamışlardır.

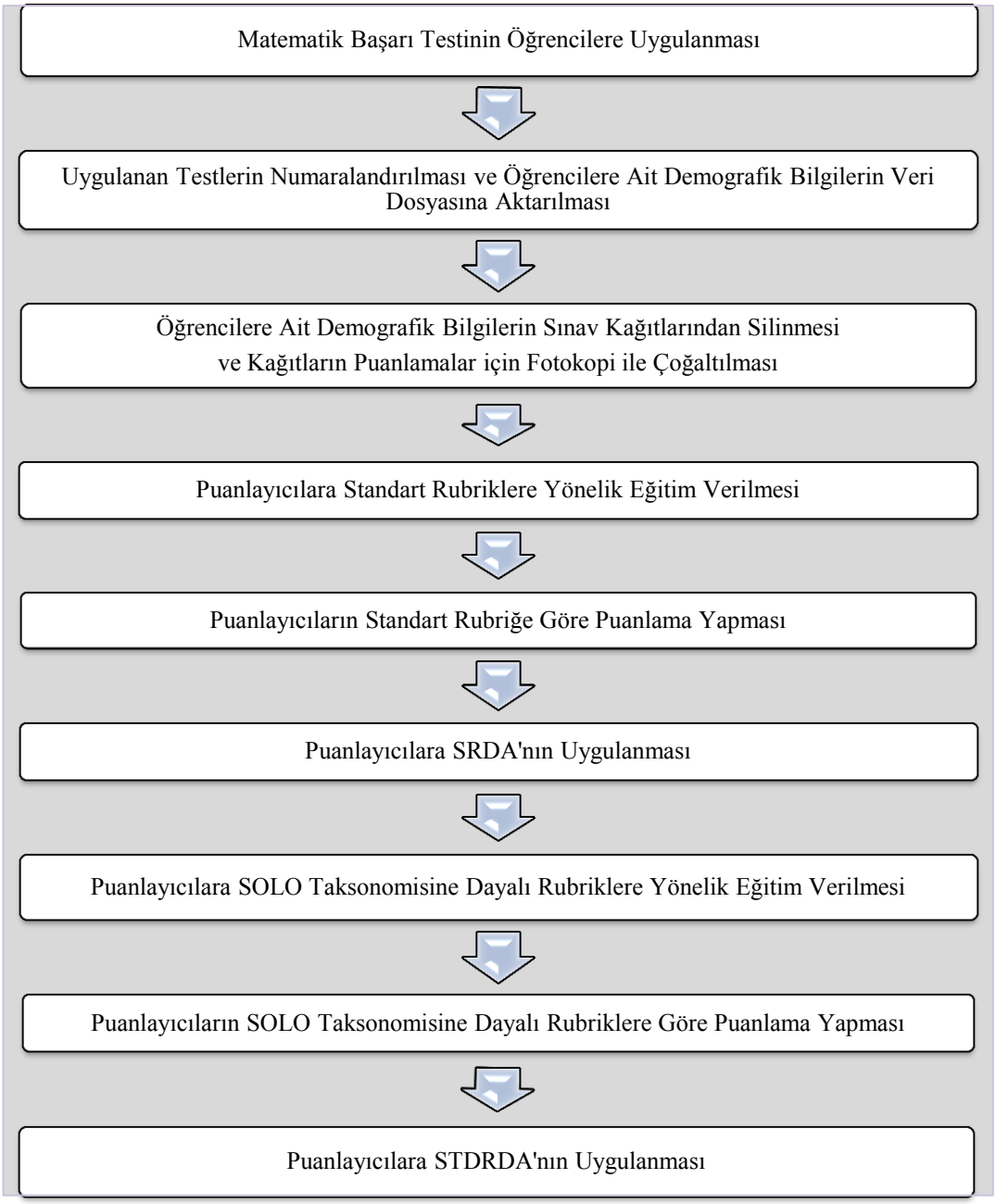
Öğrencilerden veriler toplandıktan sonra, kâğıtlar numaralandırılmış ve bu numaralara göre öğrencilerin demografik özelliklerinin yer aldığı bir veri dosyası oluşturulmuştur. Öğrencilere ait demografik özelliklerin bilgisayar ortamına aktarılmasının ardından kâğıtlar üzerindeki demografik bilgiler silinmiştir. Bu şekildeki bir uygulama ile puanlayıcıların sınav kâğıtlarını değerlendirirken, öğrencinin cinsiyeti, birinci dönem matematik karne notu ve birinci dönem yapılan matematik dersi ortak sınavındaki doğru sayısı gibi özelliklerden etkilenmesinin önüne geçilmesi amaçlanmıştır. Bu işlemin ardından kâğıtlar fotokopi ile çoğaltılmıştır. Araştırmada yedi puanlayıcının her biri, matematik başarı testini önce standart rubrik ve daha sonra SOLO taksonomisine dayalı rubrik kullanarak puanlayacağından, sınav kâğıtlarının 14 adet kopyası oluşturulmuştur. Puanlayıcılardan herhangi birinin orijinal sınav kâğıtları üzerinde puanlama yaparken, diğerlerinin fotokopi ile çoğaltılan kâğıtlar üzerinde puanlama yapması değerlendirme sonuçlarını etkileyebilecek bir faktör olarak düşünülmüştür. Puanlama işlemine bu tür bir etkinin karışmasını önlemek için puanlama sırasında sınav

kâğıtlarının orijinal hali kullanılmamıştır. Bütün puanlayıcılar değerlendirmelerini fotokopi ile çoğaltılan kâğıtlar üzerinde yapmıştır.

Öğrencilerden verilerin toplanması ve sınav kâğıtlarının puanlamalar için hazırlanması sonrasında puanlayıcılara standart rubriklere yönelik eğitimler verilmiştir. Objektif olarak puanlama yapmanın mümkün olmadığı değerlendirmelerde, puanlama işleminden önce, değerlendirilecek performansın boyutları ve bu performansı değerlendirmede kullanılacak rubriklerin kategorileri hakkında puanlayıcıların bilgilendirilmesi gerekir (Kutlu, Doğan ve Karakaya, 2010). Bu gereklilik kapsamında puanlayıcılara; *i)* performans değerlendirme, *ii)* performans değerlendirmenin avantajları ve dezavantajları, *iii)* puanlayıcı etkisi ve türleri, *iv)* puanlayıcı güvenirliği, *iv)* puanlayıcı etkilerini azaltmak için kullanılabilecek yöntemler, *vi)* rubriklerin tanımı ve geliştirilmesi ile *vi)* standart rubrikler başlıklarından oluşan bir eğitim verilmiştir. Eğitimin son basamağını puanlayıcıların standart rubrik kullanarak örnek puanlamalar yapması oluşturmuştur. Örnek puanlamalar; 13 öğrenci üzerinde yapılan ön uygulamada matematik başarı testinde yer alan ancak esas uygulamaya dâhil edilmeyen dört soru üzerinde yapılmıştır. Bu dört soru için başarılı, orta ve başarısız performans düzeylerini temsil eden öğrenci cevapları belirlenmiştir. Puanlayıcılar, eğitim sırasında yapılacak örnek uygulamalarda kullanılmak üzere geliştirilen standart rubriklerden yararlanarak öğrenci cevaplarını puanlamışlardır. Örnek uygulamaların ardından puanlayıcılara değerlendirmeleri ile ilgili dönütler verilerek standart rubriklere yönelik puanlayıcı eğitimleri tamamlanmıştır. Puanlayıcılar iki gün ile 14 gün arasında değişen sürelerde 104 öğrenciye ait kâğıtları puanlamışlardır. Puanlamaların ardından puanlayıcıların standart rubrikler hakkındaki görüşlerinin belirlenmesi amaçlanmıştır. Bu kapsamda, değerlendirmelerini tamamlayan puanlayıcılara SRDA uygulanmıştır. Anket elektronik posta ile puanlayıcılara gönderilmiştir. Anketteki maddeler puanlayıcılar tarafından yanıtlandıktan sonra yine elektronik posta ile araştırmacıya iletilmiştir.

Standart rubriğe göre yapılan puanlamaların ardından SOLO taksonomisine dayalı rubrikler kullanılarak yapılan puanlamalara geçilmiştir. Puanlayıcılar, sınav kâğıtlarını SOLO taksonomisine dayalı rubrikler ile puanlamadan önce ikinci bir puanlayıcı eğitimi gerçekleştirilmiştir. İkinci eğitim kapsamında; bir önceki eğitimde ele alınan *i)* performans değerlendirme, *ii)* performans değerlendirmenin avantajları ve dezavantajları, *iii)* puanlayıcı etkisi ve türleri, *iv)* puanlayıcı güvenirliği, *iv)*

puanlayıcı etkilerini azaltmak için kullanılabilecek yöntemler, *vi)* rubriklerin tanımı ve geliştirilmesi konuları tekrarlanmıştır. Yapılan tekrar sonrasında SOLO taksonomisi ile SOLO taksonomisine dayalı rubrik içeriklerine yer verilmiştir. İkinci puanlayıcı eğitiminin son adımında, puanlayıcılar SOLO taksonomisine dayalı rubrik kullanarak örnek puanlamalar yapmışlardır. Örnek uygulamalar birinci puanlayıcı eğitiminde olduğu gibi, 13 öğrenci üzerinde yapılan ön uygulamada matematik başarı testinde yer alan ancak esas uygulamaya dâhil edilmeyen dört soru üzerinde yapılmıştır. Söz konusu dört soru için başarılı, orta ve başarısız performans düzeylerini temsil eden öğrenci cevapları belirlenmiştir. Puanlayıcılar, eğitimde kullanılmak üzere bu sorulara yönelik olarak geliştirilen SOLO taksonomisine dayalı rubriklerden yararlanarak öğrenci cevaplarını değerlendirmişlerdir. Örnek uygulamalar sonrasında; puanlayıcılara değerlendirmeleri ile ilgili dönütler verilerek, SOLO taksonomisine dayalı rubriklere yönelik puanlayıcı eğitimleri tamamlanmıştır. Puanlayıcılar, sekiz ile 22 gün arasında değişen sürelerde puanlama işlemini tamamlamışlardır. Puanlamalar sonrasında puanlayıcıların SOLO taksonomisine dayalı rubrikler hakkında görüşlerinin belirlenmesi hedeflenmiştir. Bu doğrultuda, değerlendirmelerini tamamlayan puanlayıcılara STDRDA uygulanmıştır. Anket elektronik posta ile puanlayıcılara gönderilmiştir. Anketteki maddeler puanlayıcılar tarafından yanıtlandıktan sonra yine elektronik posta ile araştırmacıya iletilmiştir. Araştırma verilerinin toplanması sürecinde takip edilen işlemler Şekil-3.1'de ayrıca özetlenmiştir.



Şekil 3.1. Veri toplama sürecinde takip edilen işlemler

3.5. VERİ ANALİZİ

Araştırmanın verileri 104 öğrencinin açık uçlu sekiz matematik sorusuna verdiği yanıtın yedi puanlayıcı tarafından standart ve SOLO taksonomisine dayalı rubrikler ile puanlanması yoluyla elde edilmiştir. Dolayısıyla araştırmada; birey (öğrenci), madde ve puanlayıcı olmak üzere üç yüzey bulunmaktadır. Puanlayıcıların, standart ve SOLO taksonomisine dayalı rubrikleri kullanarak açık uçlu matematik sorularını puanlaması sonucunda elde edilen veriler çok yüzeyli Rasch modeline göre analiz edilmiştir. Analizler, FACET (Linacre, 2014) paket

programından yararlanılarak gerçekleştirilmiştir. Araştırmada; öğrencilere uygulanan matematik başarı testindeki sekiz sorudan altısı dörtlü derecelendirmeye sahip rubrikler ile puanlanmıştır. Diğer iki sorunun puanlanmasında kullanılan rubrikler ise beş kategorili bir yapıya sahiptir. Puanlama kategorilerinin farklı olması sebebiyle analizler karışık puanlama ölçek formuna (mixed rating scale forms) göre yürütülmüştür. Araştırmada çok yüzeyli Rasch analizleri uygulanmadan önce; söz konusu analize ilişkin varsayımların karşılanıp karşılanmadığı test edilmiştir. Bu varsayımlar arasında; tek boyutluluk, yerel bağımsızlık ve model ile veri uyumu yer almaktadır.

3.5.1. Tek Boyutluluk Varsayımı

Diğer madde tepki kuramı modellerinde olduğu gibi, Rasch modelinin bir uzantısı olan çok yüzeyli Rasch modeli için karşılanması gereken varsayımlardan biri ölçülen özelliğin tek boyutlu bir yapıya sahip olmasıdır. Bu varsayım, testte yer alan maddelerin tamamının tek bir özelliği ölçmeye yönelik olması anlamına gelmektedir (Harvey ve Hammer, 1999). Verilerin tek boyutluluk varsayımını karşılayıp karşılamadığı genellikle faktör analizi ile test edilmektedir. Dolayısıyla araştırma verilerinin tek boyutluluk varsayımını sağlayıp sağlamadığı Açımlayıcı Faktör Analizi (AFA) ile sınanmıştır. Faktör analizi; puanlayıcıların her bir maddeye verdikleri puanların ortalamaları üzerinden gerçekleştirilmiştir.

3.5.1.1. Standart Rubrik Kullanılarak Yapılan Puanlamalar için AFA Sonuçları

AFA gerçekleştirilmeden önce veri setinin faktör analizine uygun olup olmadığı incelenmiştir. Bu amaçla KMO ve Bartlett testlerinden yararlanılmaktadır. Verilerin faktör analizine uygun olabilmesi için KMO değerinin .60'tan yüksek ve Bartlett testinin istatistiksel olarak anlamlı olması gerekmektedir (Büyüköztürk, 2010). Araştırmada KMO örneklem uygunluk katsayısı .654 olarak bulunmuş ve Bartlett testinin istatistiksel olarak anlamlı olduğu (χ^2=148,644, *sd*=28; p<.01) belirlenmiştir. Bu sonuca göre, verilerin faktör analizine uygun olduğu söylenebilir. Bu tespitin ardından temel bileşenler yöntemi kullanılarak AFA gerçekleştirilmiştir. AFA sonucunda tek faktörlü bir yapı elde edileceği öngörüldüğünden, döndürme uygulanmamıştır. Standart rubrik kullanılarak yapılan puanlamalar için AFA sonuçları Tablo 3.5'te gösterilmiştir.

Tablo 3.5. Standart rubrik kullanılarak yapılan puanlamalar için AFA sonuçları

Madde No	Faktör Yükü	Madde No	Faktör Yükü	Madde No	Faktör Yükü	Madde No	Faktör Yükü
M1	.40	M3	.55	M5	.61	M7	.74
M2	.40	M4	.54	M6	.49	M8	.70
Öz Değer=2.546 ve Açıklanan Varyans=%31.823							

AFA'da toplam varyansın %31.82'sini açıklayan tek faktörlü bir yapı elde edilmiştir. Testte yer alan maddelerin faktör yüklerinin .40 ile .74 arasında değiştiği saptanmıştır. AFA'da açıklanan varyans oranı için %30 ve faktör yükü için .30 değerinin alt sınır olarak kabul edildiği bilinmektedir (Bayram, 2009). Bu ölçütlere göre, AFA'dan elde edilen bulgular; matematik başarı testinin yapı geçerliğinin sağlandığına ve tek boyutluluk varsayımının karşılandığına yönelik bir kanıt olarak değerlendirilebilir.

3.5.1.2. SOLO Taksonomisine Dayalı Rubrik Kullanılarak Yapılan Puanlamalar için AFA Sonuçları

Yukarıda belirtildiği gibi; AFA uygulanmadan önce veri setinin faktör analizine uygunluğu incelenmiştir. Yapılan incelemede, KMO değeri .722 olarak elde edilmiş ve Bartlett testi istatistiksel olarak anlamlı (χ^2=105,121, *sd*=28, p<.01) bulunmuştur. Buna göre, verilerin faktör analizine uygun olduğu söylenebilir. Döngüsüz metot ve temel bileşenler analizi kullanılarak gerçekleştirilen AFA'da başarı testindeki maddelerin toplam varyansın %30.84'ünü açıklayan tek faktör altında toplandığı saptanmıştır. SOLO taksonomisine dayalı rubrik kullanılarak yapılan puanlamalar için AFA sonuçları Tablo 3.6'da yer almaktadır.

Tablo 3.6. SOLO taksonomisine dayalı rubrik kullanılarak yapılan puanlamalar için AFA sonuçları

Madde No	Faktör Yükü	Madde No	Faktör Yükü	Madde No	Faktör Yükü	Madde No	Faktör Yükü
M1	.40	M3	.59	M5	.70	M7	.69
M2	.35	M4	.42	M6	.46	M8	.69
Öz Değer=2.467 ve Açıklanan Varyans=%30.841							

AFA'dan elde edilen bulgulara göre, testi oluşturan maddelerin faktör yüklerinin .35 ile .70 arasında sıralandığı belirlenmiştir. Bu bulgular; AFA sonuçları yorumlanırken dikkate alınması önerilen ölçütleri karşılamaktadır. Dolayısıyla; matematik başarı testinin yapı geçerliğinin sağlandığı ve tek boyutluluk varsayımının karşılandığı söylenebilir.

3.5.2. Yerel Bağımsızlık

Yerel bağımsızlık tek boyutluluk ile paralel çalışan bir varsayımdır. Dolayısıyla, tek boyutluluk varsayımının karşılandığı durumlarda yerel bağımsızlık varsayımın da karşılanmış olacağı (Hambleton, Swaminathan ve Rogers, 1991) ifade edilmektedir. Bu noktadan hareketle, araştırma verilerinin yerel bağımsızlık varsayımını karşıladığına kanaat getirilmiştir. Diğer bir ifadeyle, yerel bağımsızlık varsayımı ayrıca test edilmemiş; tek boyutluluk varsayımı karşılandığından yerel bağımsızlık varsayımının da sağlandığı kabul edilmiştir.

3.5.3. Model ile Veri Uyumu

Model ile veri arasındaki uyum, çok yüzeyli Rasch analizinden elde edilen standartlaştırılmış artık değerleri (StRes) incelenerek belirlenmektedir. Linacre (2014), model ile verilerin uyumlu olabilmesi için ±2 aralığının dışında kalan standartlaştırılmış artıkların sayısının toplam veri sayısının yaklaşık %5'inden fazla olmaması gerektiğini belirtmektedir. Yine Linacre'ye (2014) göre, model ile verinin uyumlu olabilmesi için ±3 aralığının dışında yer alan standartlaştırılmış artıkların sayısı toplam veri sayısının yaklaşık %1'ini aşmamalıdır. Araştırmada 104 öğrencinin sekiz soruya verdikleri yanıt yedi puanlayıcı tarafından puanlanmıştır. Dolayısıyla hem standart hem de SOLO taksonomisine dayalı rubrik kullanılarak yapılan değerlendirmelerde toplam 5824 (104×8×7) veri bulunmaktadır. Standart rubriğe göre yapılan puanlamalarda ±2 aralığının dışında yer alan standart artıkların sayısı 271 (%4.65) ve ±3 aralığının dışında yer alan standart artıkların sayısı 56 (%0.96) olarak bulunmuştur. Dolayısıyla standart rubriğe göre yapılan puanlamalarda, model ile veri arasındaki uyumun yeterli olduğu söylenebilir. SOLO taksonomisine göre yapılan puanlamalar incelendiğinde ise, ±2 aralığının dışında yer alan standart artıkların sayısı 289 (%4.96) ve ±3 aralığının dışında yer alan standart artıkların sayısı 91 (%1.56) olarak belirlenmiştir. Buna göre, ±3 aralığının dışında yer alan standart artıkların yüzdesinin Linacre (2014) tarafından önerilen %1 ölçütünü aştığı söylenebilir. Ancak; Linacre'nin (2014) model ile veri arasındaki uyum hakkında verilecek kararlarda dikkate alınmasını önerdiği bu ölçütleri kesin bir biçimde tanımlamayıp yaklaşık değerler olarak ifade ettiği bilinmektedir. Çok yüzeyli Rasch analizinde ±3 aralığının dışında kaldığı tespit edilen standart artıkların yüzdesi bu doğrultuda ele alındığına, model ile veri arasındaki uyumun kabul edilebilir olduğu düşünülmektedir. Nitekim McNamara (1996), ±2 ya da ±3

aralığının dışında kalan standart artıkların yüzdesi, ölçüt olarak alınması önerilen değerlerden önemli bir sapma göstermediği sürece çok yüzeyli Rasch modelinin kullanılmasından vazgeçilmemesi gerektiğini belirtmektedir. Çünkü temel madde tepki kuramında bir, iki ve üç parametreli modellerden hangisi veriler ile daha iyi uyum gösteriyorsa, analizler veri seti ile daha iyi uyum gösterdiği belirlenen modele göre yürütülmektedir. Yani üç parametreli model veriler ile yeterli uyum vermediği takdirde, iki parametreli model kullanılabilmekte ya da iki parametreli modelin veri ile uyumunun düşük olması halinde bir parametreli modelden yararlanılabilmektedir. Oysa çok yüzeyli Rasch analizinde model ile veri arasındaki uyumun yeterince yüksek olmaması durumunda, bu modelin yerine kullanılabilecek alternatif bir model bulunmamaktadır. Buna bağlı olarak, model ile veri arasındaki uyum yüksek olmasa da; performans değerlendirmede çok yüzeyli Rasch modelinin kullanılması önerilmektedir (McNamara, 1996). Dolayısıyla, SOLO taksonomisine dayalı rubrik kullanılarak yapılan puanlamalarda ±3 aralığının dışında kalan standart artıkların yüzdesinin çok yüzeyli Rasch modelinin kullanımına engel teşkil etmeyecek büyüklükte olduğu söylenebilir.

Varsayımların karşılandığı belirlendikten sonra, çok yüzeyli Rasch analizi gerçekleştirilmiştir. Ardından, analiz çıktıları her bir puanlayıcı etkisine ilişkin literatürde yer alan ölçütler doğrultusunda incelenmiştir. Puanlayıcı katılığı ve cömertliği, merkeze yönelme etkisi, halo etkisi, tutarsızlık ve yanlılık etkilerini belirlemeye yönelik olarak incelenen grup düzeyindeki ve bireysel düzeydeki istatistiksel göstergeler (Myford ve Wolfe, 2004) Tablo 3.7'de sunulmuştur.

Tablo 3.7. Puanlayıcı etkilerini belirlemeye yönelik olarak incelenen istatistiksel göstergeler

Puanlayıcı Etkileri	İstatistiksel Göstergeleri	
	Grup Düzeyinde	Bireysel Düzeyde
Puanlayıcı Katılığı ve Cömertliği	- Puanlayıcı yüzeyine ilişkin ayırma oranı, ayırma indeksi ve güvenirlik indeksinin yüksek olması - Puanlayıcı yüzeyi için hesaplanan Ki Kare değerinin istatistiksel açıdan anlamlı olması	- Puanlayıcılardan herhangi birinin değişken haritası üzerinde diğer puanlayıcılara göre daha farklı bir konumda bulunması - Puanlayıcılara ait logit ölçüleri, bu logit ölçülerinin ortalaması ve standart hatası kullanılarak hesaplanan *t* değerinin puanlayıcılardan herhangi biri için istatistiksel açıdan anlamlı olması
Merkeze Yönelme Etkisi	- Birey yüzeyine ilişkin ayırma oranı, ayırma indeksi ve güvenirlik indeksinin düşük olması - Birey yüzeyi için hesaplanan Ki Kare değerinin anlamlı olmaması - Madde yüzeyine ilişkin uygunluk istatistiklerinin kabul edilebilir alt sınır olan .5'in altında yer alması - Grup düzeyindeki kategori istatistikleri tablosuna göre, puanlama ölçeğinin orta kategorisinin diğer kategorilere kıyasla daha yoğun bir biçimde kullanılmış olması	- Puanlayıcı yüzeyine ilişkin uygunluk istatistiklerinin kabul edilebilir alt sınır olan .5'in altında yer alması - Bireysel düzeydeki kategori istatistikleri tablosunda, puanlama ölçeğinin orta kategorilerini diğer kategorilere kıyasla daha yoğun bir biçimde kullanan puanlayıcıların bulunması
Halo Etkisi	- Madde yüzeyine ilişkin ayırma oranı, ayırma indeksi ve güvenirlik indeksinin düşük olması - Madde yüzeyi için hesaplanan Ki Kare değerinin istatistiksel açıdan anlamlı olmaması	- Maddelerin güçlük düzeyleri açısından eşitlendiği modelde veri ile mükemmel uyum gösteren puanlayıcıların olması - Maddelerin güçlük düzeyleri arasındaki farklılaşma az olduğunda, herhangi bir puanlayıcıya ait uygunluk istatistiklerinin 1'den önemli ölçüde düşük olması; maddelerin güçlük düzeyleri arasındaki değişkenlik fazla olduğunda ise, herhangi bir puanlayıcıya ait uygunluk istatistiklerinin 1'den önemli ölçüde büyük olması
Tutarsızlık	- Birey yüzeyine ilişkin ayırma oranı, ayırma indeksi ve güvenirlik indeksinin düşük olması - Birey yüzeyi için hesaplanan Ki Kare değerinin anlamlı olmaması	- Birey yüzeyinin bileşenlerinin yetenek düzeyi açısından eşitlendiği modelde, veri ile mükemmel uyum gösteren puanlayıcıların olması - Herhangi bir puanlayıcının yaptığı puanlamalar ile geriye kalan diğer puanlayıcıların yaptığı puanlamalar arasındaki korelasyonun (single rater-rest of rater correlation) düşük olması
Yanlılık	- Yüzeyler arasında manidar bir etkileşim olup olmadığını gösteren Ki Kare değerinin istatistiksel açıdan anlamlı olması	- Aralarında, yanlılık olup olmadığı araştırılan yüzeylerin bileşenleri arasındaki anlamlı etkileşimler

Standart ve SOLO taksonomisine dayalı rubrikler kullanılarak yapılan puanlamalar çok yüzeyli Rasch modeline göre analiz edildikten sonra, elde edilen yetenek kestirimleri ile öğrencilerin matematik karne notları ve matematik dersi ortak sınavındaki doğru sayıları arasındaki ilişki Pearson Momentler Çarpımı korelasyonu kullanılarak incelenmiştir. Standart ve SOLO taksonomisine dayalı rubriklere göre hesaplanan yetenek kestirimleri arasındaki farkın anlamlılığı ise ilişkili örneklem *t*-testi ile sınanmıştır. İlişkili örneklem *t*-testinin ardından, tespit edilen anlamlı farkın büyüklüğüne karar vermek için etki değeri hesaplanmıştır. Etki değerinin hesaplanmasında kullanılan birçok farklı istatistik bulunmaktadır. Bununla birlikte, en yaygın olarak kullanılan istatistik Cohen'in Eta Kare sınıflamasıdır. Bu sınıflamaya göre, hesaplanan Eta kare değeri .01 ile .06 arasında ise anlamlı farkın küçük, .06 ile .14 arasında ise anlamlı farkın orta düzeyde, .14 ve üzerinde ise anlamlı farkın büyük olduğu yorumları yapılmaktadır (Pallant, 2005). Korelasyon analizi ve ilişkili örneklem *t*-testi için SPSS 20.0 paket programından yararlanılmıştır.

Araştırmada, puanlayıcıların SRDA ile STDRDA'da yer alan kapalı uçlu maddelere verdikleri yanıtların analizinde aritmetik ortalama değerlerinden faydalanılmıştır. Açık uçlu maddelere verilen yanıtlar ise betimsel analiz yaklaşımıyla çözümlenmiştir. Betimsel analiz yaklaşımında, veriler önceden belirlenen temalara göre yorumlanmaktadır (Yıldırım ve Şimşek, 2011). Bu çalışmada, açık uçlu maddelerden elde edilen veriler yorumlanırken SRDA ile STDRDA'da yer alan maddelerin ortaya koyduğu temalar dikkate alınmıştır. Dolayısıyla, açık uçlu maddelere verilen yanıtların çözümlenmesinde, betimsel analiz yaklaşımından faydalanılması uygun görülmüştür. Puanlayıcıların standart ve SOLO taksonomisine dayalı rubrikler hakkındaki görüşlerini çarpıcı bir biçimde yansıtmak amacıyla doğrudan alıntılara yer verilmiştir. Doğrudan alıntılar sunulurken puanlayıcı için P kısaltması kullanılmıştır. Ayrıca her puanlayıcı bir sayı ile kodlanmıştır. Örneğin, Puanlayıcı 1 için (P1) ve Puanlayıcı 5 için (P5) şeklinde bir kodlama yapılmıştır. Doğrudan alıntılarda, durumu en iyi yansıtan ifadelere daha çok yer verilmiştir.

DÖRDÜNCÜ BÖLÜM

BULGULAR VE TARTIŞMA

Bu bölümde, ilk olarak araştırmadan ulaşılan bulgulara yer verilmiştir. Daha sonra, elde edilen bulgular ilgili literatür dikkate alınarak tartışılmıştır.

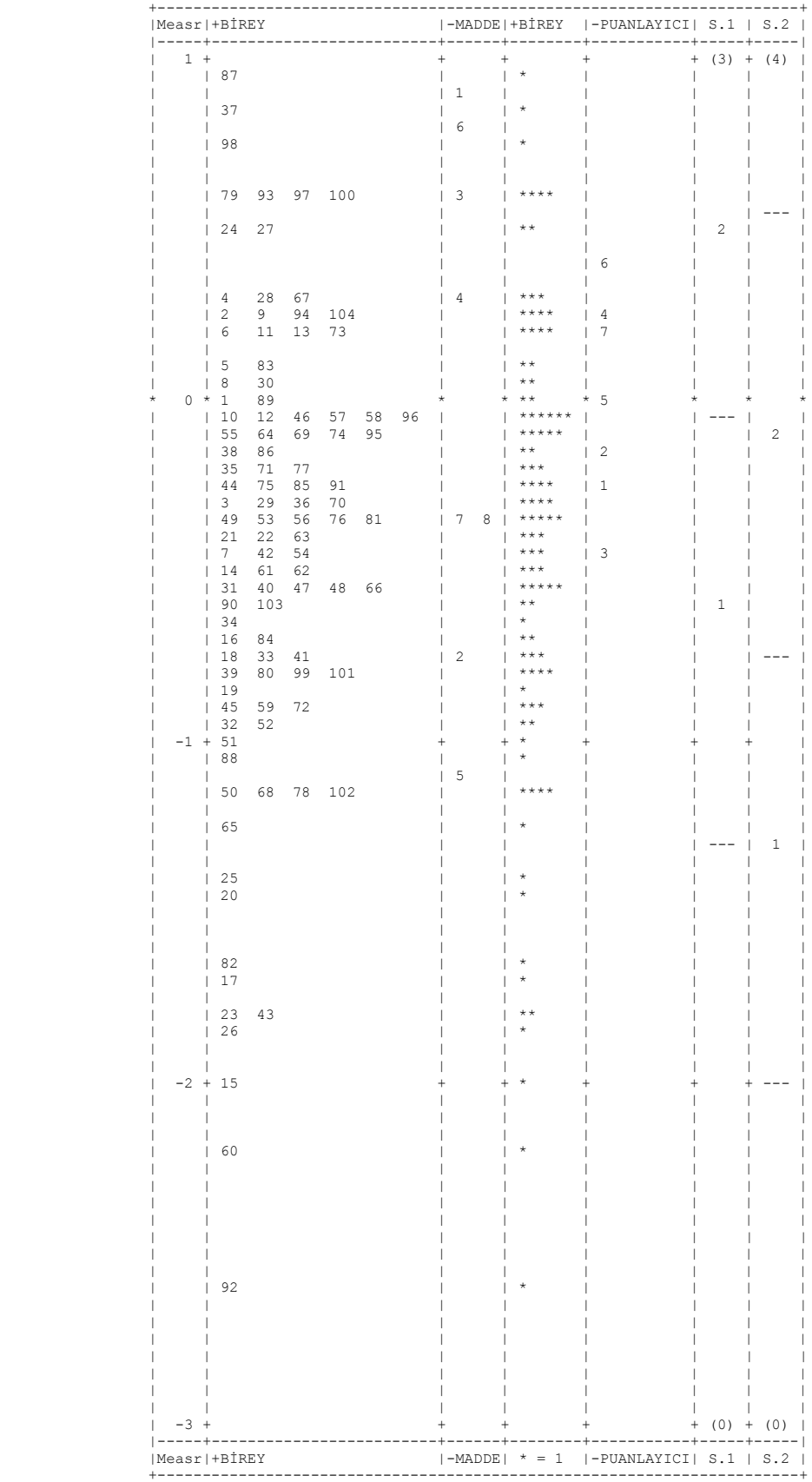
4.1. BULGULAR

Araştırmada ulaşılan bulgular, araştırma problemlerinin sırasına uygun olarak aşağıda sunulmuştur.

4.1.1. Birinci Alt Probleme İlişkin Bulgular

Araştırmanın birinci alt problemi "Standart rubriklere göre yapılan puanlamalarda, puanlayıcı, birey (öğrenci) ve madde yüzeyleri için hesaplanan güvenirlik değerleri ile uygunluk istatistikleri nasıldır?" şeklinde ifade edilmiştir. Bu alt problem doğrultusunda, öğrencilerin açık uçlu matematik sorularına verdikleri yanıtlar yedi puanlayıcı tarafından puanlanmış, puanlama sonucunda elde edilen veriler çok yüzeyli Rasch modeline göre analiz edilmiştir.

Analiz çıktılarında ilk olarak iterasyon raporu sunulmuştur. İterasyon raporu, verilerin program tarafından 17 defa okunduğunu göstermiştir. Uygulanan iterasyon sayısının az olması veriden iyi tahminde bulunmanın kolay olduğu anlamına gelmektedir. Analiz çıktılarında, iterasyon raporundan sonra değişken haritası (variable map) yer almaktadır. Standart rubrik kullanılarak yapılan puanlamaların analiz edilmesiyle elde edilen değişken haritası Şekil 4.1'de sunulmuştur.



Şekil 4.1. Standart rubriklere göre yapılan puanlamaların çok yüzeyli Rasch modeline göre analiz edilmesiyle elde edilen değişken haritası

Şekil 4.1'e bakıldığında; en yüksek yetenek düzeyindeki öğrencinin 87 numaralı öğrenci (.95 logit birimde), en düşük yetenek düzeyindeki öğrencinin ise 92 numaralı öğrenci (-2.62 logit birimde) olduğu görülmektedir. Değişken haritasının üçüncü sütununda maddeler bulunmaktadır. Maddelerin güçlük düzeyleri açısından sıralandığı bu sütunda, aşağıdan yukarı doğru gidildikçe madde güçlüğü artmaktadır. Buna göre, en zor sorunun bir numaralı madde (.90 logit biriminde), en kolay sorunun ise beş numaralı madde (-1.11 logit biriminde) olduğu belirlenmiştir. Değişken haritasının beşinci sütununda puanlayıcılara ilişkin ölçümler yer almaktadır. Puanlayıcı yüzeyi için yapılacak yorumlar, sütunun üst ucunda yer alan ve yüksek logit puanına sahip olan puanlayıcıların daha katı; sütunun alt ucunda yer alan ve düşük logit puanına sahip olan puanlayıcıların ise daha cömert puanlamalar yaptığı şeklindedir. Dolayısıyla, en katı puanlamaların altı numaralı puanlayıcı (.39 logit), en cömert puanlamaların ise üç numaralı puanlayıcı (-.45 logit) tarafından yapıldığı ortaya çıkmaktadır. Birey, madde ve puanlayıcı yüzeylerinde logit ölçeğinin negatif ve pozitif ucu boyunca uzanan ölçümlerin elde edilmesi; bireylerin yetenek düzeyleri açısından ayırt edilebildiğine, maddelerin güçlük düzeyleri açısından farklılık gösterdiğine ve puanlayıcılar arasında katılık/cömertlikleri yönüyle fark bulunduğuna işaret etmektedir. Ancak bu konuda daha kesin bir kanıya varabilmek için her bir yüzeye ait ölçüm raporları incelenmelidir.

Puanlayıcı yüzeyine ilişkin ölçüm raporları Tablo 4.1'de gösterilmiştir. Tablo 4.1'in ikinci sütununda puanlayıcıların katılık ve cömertliklerine ilişkin ölçümler yer almaktadır. Puanlayıcılar için rapor edilen pozitif logit ölçümleri puanlayıcının katı puanlamalar yaptığına işaret ederken; negatif logit ölçümleri puanlayıcının cömert puanlamalar yaptığını göstermektedir. Tablo 4.1'e göre, puanlayıcılara ilişkin logit ölçüleri .39 ile -.45 arasında değişmekte olup, puanlayıcıların katılık ve cömertliklerine ilişkin aralık .84 logittir [.39-(-.45)]. Tablo 4.1'de yer alan istatistiklerden biri de, uygunluk içi ve uygunluk dışı istatistikleridir. Uygunluk istatistiklerinin ortalamasının 1'e eşit olması, model ile veri arasındaki uyumun mükemmel olduğunu göstermektedir. Ancak, gerçek ölçme durumlarında model ile veri arasındaki uyumun mükemmel olması genellikle imkânsızdır (Brentari ve Golia, 2008). Dolayısıyla uygunluk istatistiklerine ilişkin kabul edilebilir aralığın ne olduğu sorusunun yanıtlanması gerekmektedir. Wright ve Linacre (1994), .6 ile 1.4 arasında kalan uygunluk değerlerini kabul edilebilir olarak belirtmiştir. Bu ölçüte göre, .5 ve altındaki değerler ile 1.5 ve üzerindeki değerler verilerin ölçüm için uygun olmadığı

şeklinde yorumlanmaktadır. Myford ve Wolfe (2003) ise, 2'ye kadar olan uygunluk istatistiklerini kabul edilebilir olarak nitelendirmiştir. Myford ve Wolfe'a (2003) göre, 1.5 ile 2 arasındaki değerler verilerin ölçüm için yararlı olmadığını yansıtmakta; ancak zararlı da olmadığı anlamına gelmektedir. 2'nin üstündeki uygunluk istatistikleri ise, verilerin ölçüm için zararlı olduğunu göstermektedir (Sudweeks, Reeve ve Bradshaw, 2004). Puanlayıcılar için rapor edilen uygunluk içi ve uygunluk dışı istatistiklerinin ortalamasına bakıldığında, .99 ve 1.05 gibi 1'e oldukça yakın değerler olduğu belirlenmiştir. Bu değerler verinin model ile uyumlu olduğunu göstermektedir. Ayrıca, uygunluk istatistiklerinin puanlayıcıların hiçbirinde kabul edilebilir aralığın dışında kalmadığı saptanmıştır. Bu bulgu, model ile veri uyumunu olumsuz etkileyen puanlayıcı bulunmadığını yansıtmaktadır.

Tablo 4.1. Standart rubriklere göre yapılan puanlamaların çok yüzeyli Rasch modeline göre analiz edilmesiyle puanlayıcı yüzeyi için elde edilen ölçüm raporları

Puanlayıcı	Logit Ölçüsü	Standart Hata	Uygunluk İçi	Uygunluk Dışı
P6	.39	.04	.98	1.08
P4	.26	.04	1.05	1.09
P7	.19	.04	.76	.90
P5	.02	.04	.97	1.01
P2	-.16	.04	.87	.97
P1	-.24	.04	1.16	1.21
P3	-.45	.04	1.12	1.12
Ortalama	.00	.04	.99	1.05
Standart Sapma (Evren)	.28	.00	.13	.10
Standart Sapma (Örneklem)	.30	.00	.14	.10
Model, Evren: RMSE=.04	Standart Sapma=.27	Ayırma Oranı=6.53	Güvenirlik= 98	
Model, Örneklem: RMSE=.04	Standart Sapma=.30	Ayırma Oranı=7.07	Güvenirlik=.98	
Model, Tamamı Aynı Ki Kare=306.0 sd=6 p=.00 Model, Rastgele Normal Ki Kare=5.9 sd=5 p=.32				
Puanlayıcılar arası mutlak uyum: %55.2				
Puanlayıcılar arası beklenen uyum: %38.8				
Puanlayıcılar Arası Güvenirliğe İlişkin Kappa İstatistiği: .27				

Puanlayıcı yüzeyine ilişkin ayırma oranı ve güvenirlik indeksine bakıldığında, evren ve örneklem şeklinde iki farklı model bulunduğu görülmektedir. Linacre'ye (2014) göre, herhangi bir yüzeyin olası bütün bileşenleri model içerisinde yer alıyorsa "*model, evren*" satırındaki ayırma oranı ve güvenirlik indeksi dikkate alınmalıdır. Örneğin; cinsiyet değişkenin analize dâhil edilen bir yüzey olması

halinde, kız/erkek şeklinde bu yüzeyin olası bütün bileşenleri model içerisinde yer alacaktır. Böyle bir durumda, "*model, evren*" satırında yer alan ayırma oranı ve güvenirlik indeksine bakılmaktadır. Ancak yüzeyin olası bütün bileşenlerinden yalnızca tesadüfi olarak seçilen bir kısmı model içerisinde yer alıyorsa, "*model, örneklem*" satırındaki ayırma oranı ve güvenirlik indeksi esas alınmaktadır. Söz gelimi; birey, puanlayıcı ya da madde yüzeylerinin olası bütün bileşenlerinin modele dâhil edilmesi mümkün değildir. Bu yüzeyler için, birey, puanlayıcı ve madde evreninden tesadüfi olarak seçilen bileşenler modelde yer alır. Bu şekildeki bir durumda, "*model, örneklem*" satırındaki ayırma oranı ve güvenirlik indeksinin yorumlanması gerekmektedir (Linacre, 2014). Buna bağlı olarak, puanlayıcı yüzeyine ilişkin bulgular yorumlanırken; "*model, örneklem*" satırındaki ayırma oranı ile güvenirlik indeksi dikkate alınmıştır. Ayırma oranı 7.07 ve güvenirlik indeksi .98 olarak belirlenmiştir. Puanlayıcı yüzeyi için hesaplanan güvenirlik indeksi, puanlayıcılar arasındaki güvenilir benzerliği değil; güvenilir farkı göstermektedir (Haiyang, 2010). Bu güvenirliğin yüksek olması puanlayıcıların katılıkları/ cömertlikleri yönüyle farklılık gösterdiği anlamına gelmektedir. Bu araştırmada puanlayıcı yüzeyine ilişkin .98 gibi yüksek bir güvenirlik indeksinin hesaplanması, puanlayıcıların katılık/cömertlik açısından farklılık gösterdiğine işaret etmektedir. Bu farkın anlamlı olup olmadığına Ki Kare değerine bakılarak karar verilmektedir. Çok yüzeyli Rasch analizinde her bir yüzey için *rastgele normal* ve *tamamı aynı* olmak üzere iki farklı Ki Kare değeri rapor edilmektedir. Herhangi bir yüzeyin bileşenlerinin normal dağılıma sahip bir evrenden tesadüfi olarak seçilen bir örneklemi temsil edip etmediğine karar vermek için *rastgele normal* Ki Kare değeri referans alınmaktadır. Ölçüm hatasına izin verildikten sonra, yüzeyin bileşenleri arasında anlamlı fark olup olmadığını belirlemek için ise *tamamı aynı* Ki Kare incelenmektedir (Linacre, 2014). Buna göre, katılık ve cömertlikleri yönüyle puanlayıcılar arasında anlamlı fark olup olmadığını belirlemek için tamamı aynı Ki Kare değeri incelenmiştir. Ki Kare değeri istatistiksel olarak manidar [χ^2=306.00, *sd*=6, *p*<.01] olduğundan, katılık ve cömertlikleri açısından puanlayıcılar arasında anlamlı fark bulunduğu saptanmıştır.

Tablo 4.1'de puanlayıcı güvenirliğine ilişkin sunulan diğer istatistikler; puanlayıcılar arası mutlak uyum, puanlayıcılar arası beklenen uyum ve puanlayıcılar arası güvenirliğe ilişkin kappa istatistiğidir. Tablo 4.1'e göre, puanlayıcılar arası beklenen uyum %38.8'dir. Puanlayıcılar arası mutlak uyum ise %55.2 olarak

hesaplanmıştır. Mutlak uyum, puanlayıcıların aynı öğrenci cevabına ne derece aynı puanı verdiklerini göstermektedir. Literatür incelendiğinde, mutlak uyum yüzdesine ilişkin dikkate alınması gereken ölçütün ne olduğuna dair araştırmacılar arasında bir uzlaşı bulunmadığı görülmektedir. Bununla birlikte, kabul edilebilir bir puanlayıcı güvenirliğinden söz edilebilmesi için puanlayıcılar arası mutlak uyumun %75 ve üzerinde olması önerilmektedir (Graham, Milanowski ve Miller, 2012). Bu ölçüt, standart rubriklere göre yapılan puanlamalarda puanlayıcılar arası güvenirliğin düşük olduğunu göstermektedir. Tablo 4.1'de puanlayıcılar arası güvenirliğe ilişkin yer alan bir diğer istatistik Kappa değeridir. Kappa değeri çok yüzeyli Rasch analizi çıktılarında doğrudan sunulan bir istatistik değildir. Ancak, analiz çıktılarında sunulan gözlenen ve beklenen uyum değerleri kullanılarak Kappa istatistiği hesaplanabilmektedir. Kappa istatistiğinin hesaplanmasında aşağıdaki formülden yararlanılmaktadır.

$$\text{Kappa İstatistiği} = \frac{\text{Gözlenen Uyum} - \text{Beklenen Uyum}}{100 - \text{Beklenen Uyum}}$$

Rasch analizi sonucunda hesaplanan gözlenen ve beklenen uyum değerleri bu formülde yerine konulduğunda, kappa istatistiği .27 olarak bulunmuştur. Kappa istatistiği için .40'ın altındaki değerler kötü uyumu, .40 ile .59 arasındaki değerler zayıf uyumu, .60 ile .74 arasındaki değerler iyi uyumu, .75 ve üzerindeki değerler ise mükemmel uyumu yansıtmaktadır (Landis ve Knoch, 1977). Bu aralıklar, standart rubrik kullanılarak yapılan puanlamalarda puanlayıcılar arası uyumun kötü olduğunu göstermektedir.

Çok yüzeyli Rasch analizi çıktılarında puanlayıcı yüzeyine ilişkin ölçümlerden sonra birey yüzeyine ait ölçümler sunulmaktadır. Birey yüzeyine ilişkin ölçümler Tablo 4.2'de verilmiştir. Tablo 4.2'de bireylerin yetenek düzeylerine ilişkin logit ortalaması ile standart sapması, uygunluk istatistiklerinin ortalaması, ayırma oranı ile güvenirlik indeksi, tamamı aynı ve rastgele normal Ki Kare değerleri yer almaktadır. Birey yüzeyinde 104 bileşen bulunduğundan, Tablo 4.2'de öğrencilerin her birine ilişkin ölçümlere yer verilmesi mümkün olmamıştır.

Tablo 4.2. Standart rubriklere göre yapılan puanlamaların çok yüzeyli Rasch modeline göre analiz edilmesiyle birey yüzeyi için elde edilen ölçüm raporları

	Logit Ölçüsü	Standart Hata	Uygunluk İçi	Uygunluk Dışı
Ortalama	-.43	.16	1.00	1.05
Standart Sapma (Evren)	.67	.05	.47	.68
Standart Sapma (Örneklem)	.67	.05	.48	.68
Model, Evren: RMSE=.16	Standart Sapma =.65	Ayırma Oranı=3.96	Güvenirlik=.94	
Model, Örneklem: RMSE=.05	Standart Sapma=.65	Ayırma Oranı=3.98	Güvenirlik=.94	
Model, Tamamı Aynı Ki Kare=1458.1 sd=103 p=.00				
Model, Rastgele Normal Ki Kare=95.7 sd=102 p=.66				

Birey yüzeyine ilişkin ölçüm raporları incelendiğinde, öğrencilerin yetenek düzeylerine ilişkin kestirimlerin .95 logit ile -2.62 logit arasında değiştiği belirlenmiştir. Bu değerler, öğrencilerin yetenek düzeylerinin 3.57 logit [.95-(-2.62)] gibi geniş bir aralıkta değiştiğini göstermektedir. Bireylerin yetenek düzeylerinin ortalaması -.43 ve standart sapması .67 logit olarak hesaplanmıştır. Uygunluk içi ve uygunluk dışı istatistiklerinin ortalaması sırasıyla 1.00 ve 1.05 şeklindedir. Uygunluk istatistiklerinin 1'e eşit olması halinde, model ile veri arasındaki uyumun mükemmel olduğu bilinmektedir (Brentari ve Golia, 2008). Dolayısıyla, birey yüzeyi için elde edilen uygunluk içi ve uygunluk dışı istatistiklerinin ortalaması model ile veri arasındaki uyumun iyi olduğuna işaret etmektedir. Tablo 4.2'ye göre, ayırma oranı 3.98 ve güvenirlik indeksi .94 olarak bulunmuştur. Hesaplanan ayırma oranı ve güvenirlik indeksinin yüksek olması, yetenek düzeyi farklı olan öğrencilerin birbirinden başarılı bir biçimde ayırt edilebildiğine işaret etmektedir. Öğrencilerin yetenek düzeyleri arasındaki değişkenliklerin analiz sonuçlarına anlamlı bir fark olarak yansıyıp yansımadığı Ki Kare değeri incelenerek belirlenmektedir. Hesaplanan Ki Kare değeri [χ^2=1458.1, *sd*=103, p<.01], yetenek düzeyleri açısından öğrenciler arasında anlamlı fark bulunduğunu ortaya koymaktadır. Dolayısıyla, standart rubrik kullanılarak yapılan puanlamalarda yetenek düzeyleri farklı olan öğrencilerin birbirinden ayırt edilebildiği söylenebilir.

Madde yüzeyine ilişkin ölçüm raporları ise, Tablo 4.3'te gösterilmiştir. Tablo 4.3'e göre, maddelerin güçlük düzeyleri .90 logit ile -1.11 logit arasında değişmektedir. Buna göre, maddelerin güçlük düzeyleri arasında 2.01 [.90-(-1.11)] logitlik bir değişim gözlenmektedir. Maddelerin logit değerlerinin ortalaması .00 ve standart sapması .75 logit olarak bulunmuştur. Maddelere ilişkin uygunluk içi ve

uygunluk dışı istatistiklerinin ortalaması sırasıyla .99 ve 1.05 şeklindedir. Bu değerler, uygunluk istatistiklerinin beklenen değeri olan 1'e oldukça yakındır. Dolayısıyla, verinin model ile iyi uyum gösterdiği saptanmıştır. Ayrıca testte, Myford ve Wolfe (2003) tarafından uygunluk istatistiklerine ilişkin dikkate alınması önerilen .05 ile 2 aralığının dışında kalan madde olmadığı belirlenmiştir. Buna göre, testte model ile veri arasındaki uyumu olumsuz yönde etkileyen bir soru bulunmadığı ifade edilebilir.

Tablo 4.3. Standart rubriklere göre yapılan puanlamaların çok yüzeyli Rasch modeline göre analiz edilmesiyle madde yüzeyi için elde edilen ölçüm raporları

Madde	Logit Ölçüsü	Standart Hata	Uygunluk İçi	Uygunluk Dışı
1	.90	.05	1.24	1.32
6	.78	.05	.93	.89
3	.58	.05	1.01	1.04
4	.31	.04	.85	1.17
7	-.34	.04	.70	.70
8	-.35	.04	1.12	1.12
2	-.77	.04	1.19	1.31
5	-1.11	.05	.91	.90
Ortalama	.00	.05	.99	1.05
Standart Sapma (Evren)	.70	.00	.17	.20
Standart Sapma (Örneklem)	.75	.00	.18	.22
Model, Evren: RMSE=.05	Standart Sapma=.70		Ayırma Oranı=15.43	Güvenirlik=1.00
Model, Örneklem: RMSE=.05	Standart Sapma=.75		Ayırma Oranı=16.50	Güvenirlik=1.00
Model, Tamamı Aynı Ki Kare=1822.2 sd=7 p=.00 Model, Rastgele Normal Ki Kare=7.0 sd=6 p=.32				

Tablo 4.3'e göre, madde yüzeyine ilişkin ayırma oranı 16.50 ve güvenirlik indeksi 1.00 olarak elde edilmiştir. Madde yüzeyi için hesaplanan ayırma oranı ve güvenirlik indeksinin yüksek olması, testteki maddelerin güçlük düzeyleri açısından farklılık gösterdiğine işaret etmektedir. Ayırma oranı ile güvenirlik indeksinin işaret ettiği bu farkın anlamlı olup olmadığını belirlemek için Ki Kare değerine bakılmıştır. Ki-Kare değerinin anlamlı çıkması [χ^2=1822.22, *sd*=7, *p*<.01], güçlük düzeyleri açısından maddeler arasında manidar bir fark bulunduğunu göstermektedir.

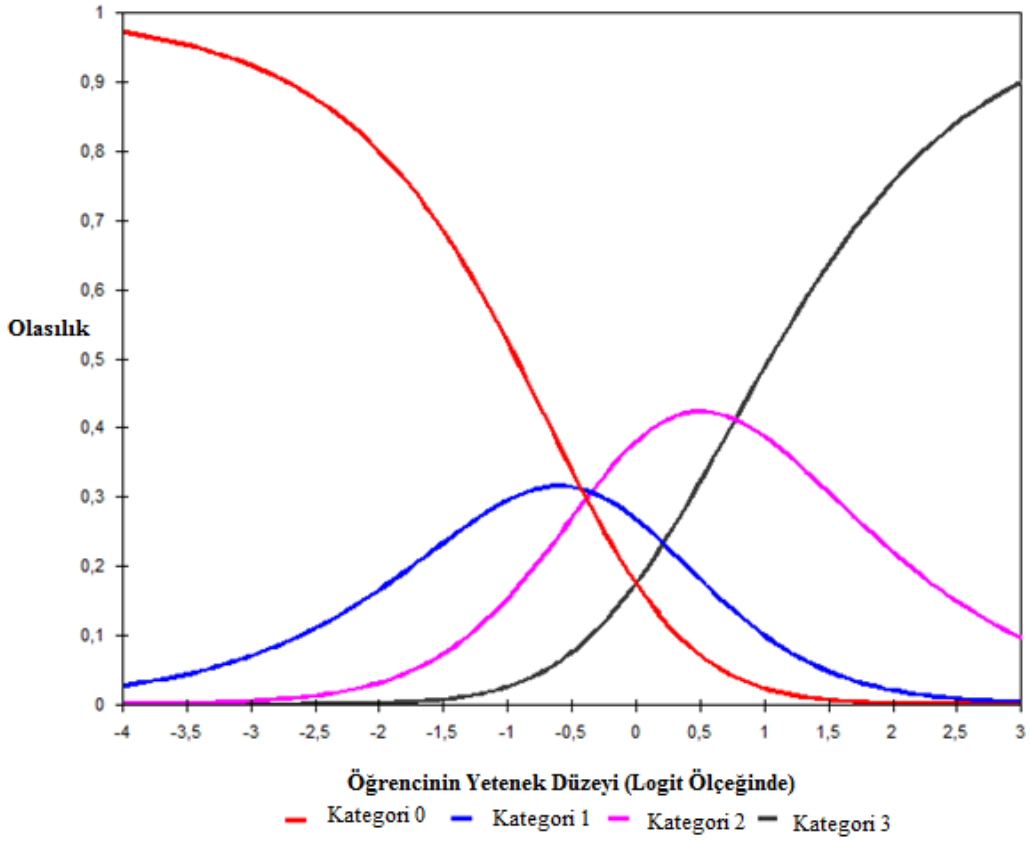
Çok yüzeyli Rasch analizi çıktılarında her bir yüzeye ilişkin ölçüm raporlarından sonra kategori istatistikleri rapor edilmektedir. Standart rubriklere göre yapılan puanlamalardan elde edilen kategori istatistikleri Tablo 4.4'te sunulmuştur. Başarı testindeki ilk altı sorunun puanlanmasında kullanılan standart rubrikler dörtlü

bir derecelendirmeye sahip iken; yedi ve sekiz numaralı soruların değerlendirilmesinde kullanılan rubrikler beşli bir derecelendirmeye sahiptir. Tablo 4.4'te, dörtlü ve beşli derecelendirmeye sahip rubriklere ilişkin kategori istatistikleri bir arada verilmiştir.

Tablo 4.4. Standart rubriklere göre yapılan puanlamaların çok yüzeyli Rasch modeline göre analiz edilmesiyle elde edilen kategori istatistikleri

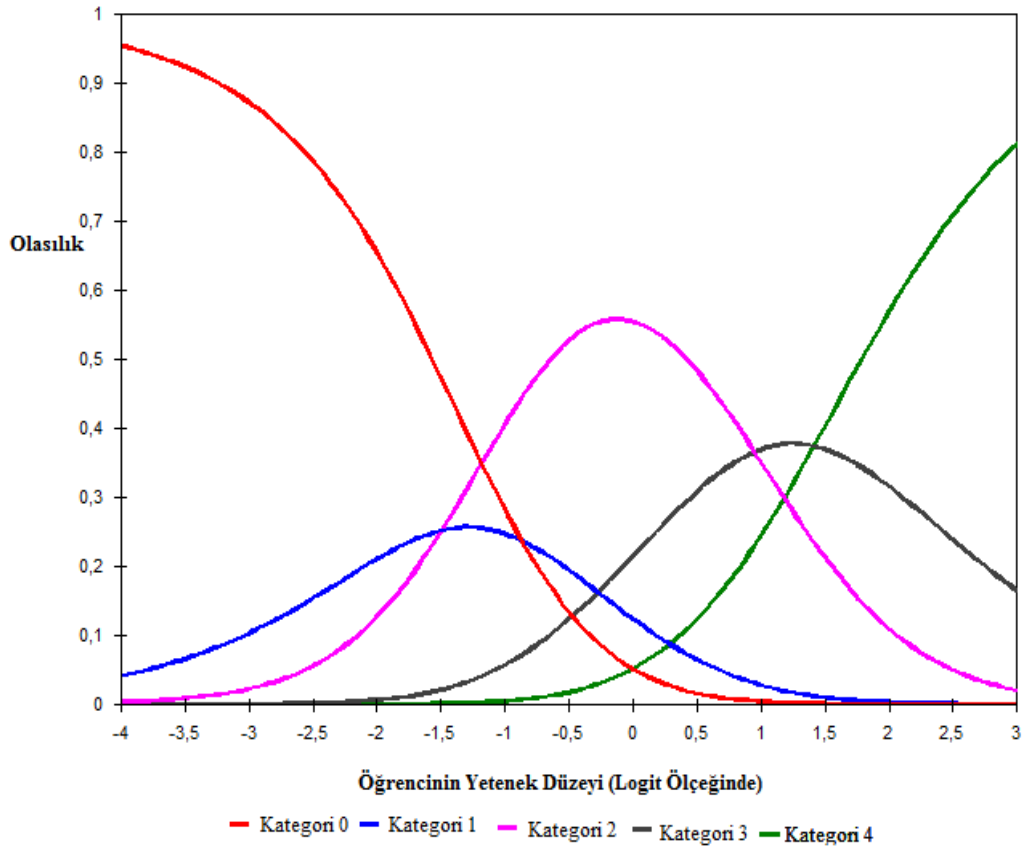
Maddeler	Puanlama Ölçeği Kategorileri	Frekans	Yüzde	Yığılmalı Yüzde	Ortalama Ölçüm	Beklenen Ölçüm	Uygunluk Dışı İstatistikleri
1-6 numaralı maddeler	0	1696	39	39	-1.24	-1.27	1.1
	1	995	23	62	-.75	-.65	.9
	2	1018	23	85	.05	.01	1.2
	3	659	15	100	.61	.65	1.1
7 ve 8 numaralı maddeler	0	161	11	11	-1.00	-.94	.9
	1	195	13	24	-.46	-.44	.9
	2	688	47	72	-.07	-.05	.9
	3	300	21	92	.35	.28	.8
	4	112	8	100	.65	.27	.9

Tablo 4.4'ün üç, dört ve beşinci sütunlarında sırasıyla; hangi puanlama kategorisinin kaç defa kullanıldığını gösteren frekans değerleri, bu frekans değerlerine ilişkin yüzdeler ve yığılmalı yüzdeler yer almaktadır. Standart rubriklerin her bir kategorisinde en az 10 gözlem bulunması, puanlama ölçeğinin etkili bir biçimde çalıştığını göstermektedir. Puanlama ölçeğinin iyi çalıştığına işaret eden bir diğer gösterge, puanlama ölçeğinin kategorileri arttıkça ortalama ölçümlerin de artmasıdır (Linacre, 2014). Tablo 4.4'teki ortalama ölçümlerin monoton olarak artması ve uygunluk dışı istatistiklerinin 1'e oldukça yakın olması puanlama ölçeğinin etkin bir biçimde çalıştığına işaret eden diğer kanıtlardır. Veriler kategorilere göre sınıflandırıldığında, uygunluk içi istatistikleri uygunluk dışı istatistiklerine yaklaşmaktadır. Dolayısıyla, kategori istatistikleri tablosunda yalnızca uygunluk dışı istatistikleri yer almakta; uygunluk içi istatistikleri ise bulunmamaktadır. Puanlama ölçeğindeki kategorilerin etkin bir biçimde çalışıp çalışmadığına karar vermek için yapılabilecek bir diğer işlem "Puanlama Ölçeği Kategorileri Olasılık Eğrisinin" incelenmesidir. Puanlama ölçeği kategorilerinin olasılık eğrisi Şekil 4.2 ve Şekil 4.3'te gösterilmiştir. Bu eğrilerin yatay ekseni yetenek düzeyini temsil etmektedir. Dikey eksen ise, olasılığı yansıtmakta olup 0 ile 1 arasında değişen değerlere sahiptir.



Şekil 4.2. Başarı testindeki ilk altı madde için standart rubrik kategorilerinin olasılık eğrisi

Şekil 4.2 incelendiğinde, düşük yetenek düzeylerinde sıfır kategorisinin, yüksek yetenek düzeylerine ise üç kategorisinin kullanılma olasılığının daha yüksek olduğu görülmektedir. Yine, Şekil 4.2'ye göre, yüksek yetenek düzeylerinde iki kategorisinin kullanılma olasılığı sıfır ve bir kategorilerinin kullanılma olasılığına göre daha fazladır. Yüksek yetenek düzeylerinde puanlama ölçeğinin üst ucundaki kategorilerin; düşük yetenek düzeylerinde ise puanlama ölçeğinin alt ucundaki kategorilerin kullanılma olasılığının daha fazla olması ilk altı maddede puanlama ölçeğinin etkin bir biçimde çalıştığına yönelik bir kanıt olarak değerlendirilebilir. Benzer şekilde Şekil 4.3'e bakıldığında, yetenek ölçeğinin alt ucundan üst ucuna doğru ilerlerken, puanlama ölçeğinin alt kategorilerinin kullanım olasılığının azaldığı; üst kategorilerinin ise kullanım olasılığının arttığı belirlenmiştir. Bu bulgu, yedi ve sekiz numaralı sorularda da puanlama ölçeğinin etkin bir biçimde çalıştığına işaret etmektedir.



Şekil 4.3. Başarı testindeki yedi ve sekiz numaralı maddeler için standart rubrik kategorilerinin olasılık eğrisi

4.1.2. İkinci Alt Probleme İlişkin Bulgular

Araştırmanın ikinci alt problemi "Standart rubriklere göre yapılan puanlamalarda, *i)* puanlayıcı katılığı ve cömertliği, *ii)* merkeze yönelme etkisi, *iii)* halo etkisi, *iv)* tutarsızlık ve *v)* yanlılık bulunmakta mıdır?" şeklinde ifade edilmiştir. Çok yüzeyli Rasch analizinden elde edilen çıktılar her bir puanlayıcı etkisine ilişkin literatürde yer alan ölçütler doğrultusunda incelenerek bu alt probleme yanıt aranmıştır. Sıralanan puanlayıcı etkilerine yönelik olarak elde edilen bulgular başlıklar halinde aşağıda sunulmuştur.

4.1.2.1. Puanlayıcı Katılığı ve Cömertliği

Puanlayıcıların katılık ve cömertlikleri yönüyle farklılık gösterip göstermediğini saptamak için ilk olarak grup düzeyindeki istatistiksel göstergeler incelenmiştir. Grup düzeyinde yapılan incelemelerde, puanlayıcı katılığı ya da cömertliğinin tespit edilmesi durumunda bireysel düzeydeki incelemelere geçilmiştir. Bireysel düzeydeki istatistiksel göstergeler yardımıyla, puanlayıcı

katılığının/cömertliğinin hangi puanlayıcı ya da puanlayıcılardan kaynaklandığı belirlenmeye çalışılmıştır.

Puanlayıcı katılığı/cömertliğine ilişkin grup düzeyinde yapılan incelemelerde; ayırma oranı, güvenirlik indeksi ve Ki Kare değerinden yararlanılmıştır. Ayırma oranı 7.07 ve güvenirlik indeksi .98 olarak bulunmuştur. Ayırma oranı ile güvenirlik indeksinin yüksek çıkması, katılık ve cömertlikleri yönüyle puanlayıcılar arasında fark bulunduğuna işaret etmektedir. Ancak, bu farkın anlamlı olup olmadığını belirleyebilmek için Ki Kare değerine bakılması gerekir. Ki Kare değerinin anlamlı çıkması [χ^2=306.0, *sd*=6, *p*<.01], katılık ve cömertlikleri yönüyle puanlayıcılar arasında gözlenen farkın manidar olduğunu göstermektedir.

Grup düzeyinde yapılan incelemelerde puanlayıcıların katılık ve cömertlikleri yönüyle farklılık gösterdiği belirlendikten sonra, bu farkın hangi puanlayıcı/puanlayıcılardan kaynaklandığının ortaya konulması gerekir. Bu doğrultuda, bireysel düzeydeki incelemelere geçilmiştir. Puanlayıcı katılığı ve cömertliğinin bireysel düzeydeki istatistiksel göstergelerinden biri değişken haritasıdır. Puanlayıcılar arasındaki farkın kaynağını, değişken haritasının puanlayıcı sütununda, diğer puanlayıcılara göre daha farklı bir noktada bulunan puanlayıcı/puanlayıcıların oluşturduğu kabul edilir (Myford ve Wolfe, 2004). Şekil 4.1'deki değişken haritasına bakıldığında, tüm puanlayıcıların logit ölçeğinin farklı noktalarında bulunduğu görülmektedir. Beş numaralı puanlayıcı logit ölçeğinin sıfır noktasında yer alırken; altı, dört ve yedi numaralı puanlayıcılar logit ölçeğinin pozitif ucunda, iki, bir ve üç numaralı puanlayıcılar logit ölçeğinin negatif ucunda yer almıştır. Puanlayıcıların değişken haritası üzerinde bulunduğu noktalar; altı, dört ve yedi numaralı puanlayıcıların puanlamada daha katı; iki, bir ve üç numaralı puanlayıcıların ise puanlamada daha cömert davrandığını düşündürmektedir. Bununla birlikte, bu konuda daha kesin bir yargıya varabilmek için her bir puanlayıcıya ilişkin *t* değerlerinin hesaplanması gerekir. *t* değeri hesaplanırken, herhangi bir puanlayıcıya ait logit ölçümünden, tüm puanlayıcılara ilişkin logit ortalaması çıkarılmakta ve elde edilen fark logit ölçümlerinin standart hatasına bölünmektedir. Örneğin; altı numaralı puanlayıcı için *t* değeri hesaplanırken; bu puanlayıcının logit ölçüsü olan .39'dan tüm puanlayıcıların logit ölçümlerinin ortalaması olan .00 çıkarılmış ve elde edilen fark puanlayıcılara ait logit ölçümlerinin standart hatası olan .04'e bölünmüştür. Böylelikle altı numaralı puanlayıcıya ait *t* değeri 9.75 olarak hesaplanmıştır. Son adımda, hesaplanan *t* değeri ile ilgili

serbestlik derecesindeki kritik *t* değeri karşılaştırılarak anlamlılık sınaması yapılmıştır. Araştırmaya yedi puanlayıcı dâhil edildiğinden, serbestlik derecesi 7-1=6 olarak hesaplanmış ve altı serbestlik derecesi ile .01 düzeyindeki kritik *t* değeri 3.71 olarak belirlenmiştir. $|t_{hesaplanan}|=9.75 > t_{kritik(.01,\ 6)} =3.71$ olduğundan, altı numaralı puanlayıcı için hesaplanan *t* değerinin anlamlı olduğu saptanmıştır. Buna göre, altı numaralı puanlayıcının puanlayıcılar arasında gözlenen farkın kaynağını oluşturan puanlayıcılardan biri olduğu söylenebilir. Bu işlem her bir puanlayıcı için tekrarlanmış olup hesaplanan *t* değerleri Tablo 4.5'te sunulmuştur. Tablo 4.5'e göre; altı, dört ve yedi numaralı puanlayıcıların puanlamada daha katı; iki, bir ve üç numaralı puanlayıcıların ise daha cömert davrandığı tespit edilmiştir.

Tablo 4.5. Standart rubrik kullanılarak yapılan puanlamalarda katılık ve cömertlikleri açısından puanlayıcılar arasında gözlenen farkın anlamlılığına ilişkin *t*-testi sonuçları

Puanlayıcı	*t* değeri	Farkın Anlamlılığı
P6	9.75	$\|t_{hesaplanan}\|>t_{kritik}$ olduğundan fark anlamlıdır. 6, 4 ve 7 numaralı puanlayıcıların her üçü de değişken haritasının pozitif ucunda yer aldığından, bu puanlayıcıların diğer puanlayıcılara kıyasla anlamlı derecede daha katı puanlamalar yaptığı belirlenmiştir.
P4	6.50	
P7	4.75	
P5	.05	$\|t_{hesaplanan}\|<t_{kritik}$ olduğundan fark anlamlı değildir.
P2	-4.00	$\|t_{hesaplanan}\|>t_{kritik}$ olduğundan fark anlamlıdır. 2, 1 ve 3 numaralı puanlayıcıların her üçü de değişken haritasının negatif ucunda yer aldığından, bu puanlayıcıların diğer puanlayıcılara göre anlamlı derecede daha cömert puanlamalar yaptığı belirlenmiştir.
P1	-6.00	
P3	-11.25	

4.1.2.2. Merkeze Yönelme Etkisi

Puanlamalarda merkeze yönelme etkisinin bulunup bulunmadığını tespit etmek için ilk olarak grup düzeyindeki istatistiksel göstergeler incelenmiş, daha sonra bireysel düzeydeki incelemelere geçilmiştir. Grup düzeyindeki incelemeler kapsamında öncelikle kategori istatistikleri tablosuna bakılmıştır (Tablo 4.4). Kategori istatistikleri tablosuna göre, testteki ilk altı madde puanlanırken rubriklerdeki tüm kategoriler dengeli bir biçimde kullanılmıştır. Puanlama ölçeğindeki kategorilerin dengeli bir biçimde kullanılmış olması, merkeze yönelme etkisinin bulunmadığına işaret etmektedir. Yedi ve sekiz numaralı maddelerde ise, puanlama ölçeğinin orta kategorisinin (kategori üç), %47 gibi bir oranla diğer kategorilere göre daha yoğun bir biçimde kullanıldığı belirlenmiştir. Bu durum, puanlamalara merkeze yönelme etkisinin karışmasının bir sonucu olabileceği gibi; öğrencilerin bu iki soruda orta düzeyde bir performans sergilemesinden de kaynaklanıyor olabilir. Dolayısıyla kategori istatistikleri tablosunun incelenmesi,

puanlamalarda merkeze yönelme etkisinin bulunup bulunmadığı konusunda karar vermek için tek başına yeterli olamamaktadır.

Puanlamalarda merkeze yönelme etkisinin bulunup bulunmadığını tespit etmek için birey yüzeyine ilişkin ayırma oranı, ayırma indeksi, güvenirlik indeksi ve Ki Kare değerlerinden de yararlanılmaktadır. Ayırma oranı, ayırma indeksi ile güvenirlik indeksinin yüksek olması ve Ki Kare değerinin anlamlı çıkması, yetenek düzeyleri farklı olan öğrencilerin birbirinden ayırt edilebildiğini göstermektedir. Öğrencilerin birbirinden başarılı bir biçimde ayırt edilebilmesi ise puanlamalarda merkeze yönelme etkisinin bulunmadığı şeklinde yorumlanmaktadır. Standart rubriklere göre yapılan puanlamalarda, birey yüzeyi için elde edilen ayırma oranı 3.98 ve güvenirlik indeksi .94 olarak bulunmuştur. Bu değerlerin yüksek olması, puanlamalarda merkeze yönelme etkisinin bulunmadığına işaret etmektedir. Birey yüzeyi için hesaplanan ayırma indeksi ise 5.64 olarak bulunmuştur. Ayırma indeksi Rasch analizi çıktılarında doğrudan sunulan bir istatistik olmayıp, ayırma oranı kullanılarak hesaplanmaktadır. Ayırma indeksinin hesaplanmasına yönelik formül aşağıda sunulmuştur.

$$\text{Ayırma İndeksi} = \frac{4 \times (\text{Ayırma Oranı}) + 1}{3}$$

Hesaplanan ayırma indeksine göre, öğrencilerin yetenekleri düzeyleri açısından yaklaşık altı katmana ayrıldığı söylenebilir. Bu bulgu, yetenek düzeyi farklı olan öğrencilerin birbirinden ayırt edilebildiğine ve puanlamalarda merkeze yönelme etkisinin bulunmadığına yönelik bir kanıt olarak değerlendirilebilir. Birey yüzeyine ilişkin Ki Kare değerinin anlamlı olması [χ^2=1458.1, *sd*=103, *p*<.01], yetenek düzeyleri açısından öğrenciler arasında gözlenen farkların manidar olduğu anlamına gelmekte ve puanlamalarda merkeze yönelme etkisinin bulunmadığını yansıtmaktadır.

Puanlamalarda merkeze yönelme etkisinin bulunup bulunmadığına karar vermek için madde yüzeyine ilişkin uygunluk istatistiklerinden de yararlanılabilmektedir. Madde yüzeyine ilişkin uygunluk istatistiklerinin, bu istatistiklerin beklenen değeri olan 1'den önemli ölçüde düşük olması, merkeze yönelme etkisine işaret etmektedir. Madde yüzeyine ilişkin ölçüm raporlarına bakıldığında, testteki maddelerden hiç birinin uygunluk içi ve uyguluk dışı istatistiklerinin kabul edilebilir alt sınırı olan .05'in altında yer almadığı saptanmıştır. Buna göre, standart rubrikler kullanılarak yapılan puanlamalarda merkeze yönelme etkisinin bulunmadığı söylenebilir.

Puanlamalarda merkeze yönelme etkisinin olup olmadığı konusunda karar vermek için grup düzeyindeki incelemelerin yanı sıra bireysel düzeydeki incelemelerden de yararlanılması gerekmektedir. Bireysel düzeydeki incelemeler kapsamında ilk olarak her bir puanlayıcıya ilişkin uygunluk istatistiklerine bakılmıştır. Uygunluk istatistiklerinin kabul edilebilir alt sınır olan .5'in altında olması, puanlayıcıların aşırı tutarlı puanlamalar yaptığını göstermektedir. Aşırı tutarlı puanlamalar ise, puanlamalara merkeze yönelme etkisinin karıştığına işaret etmektedir. Standart rubriklere göre yapılan puanlamalarda; puanlayıcılar için hesaplanan uygunluk içi istatistiklerinin .76 ile 1.12 arasında, uygunluk dışı istatistiklerinin ise .97 ile 1.21 arasında değiştiği belirlenmiştir. Bu değerler, standart rubriklere göre yapılan puanlamalarda, merkeze yönelme etkisinin bulunmadığı anlamına gelmektedir. Ancak; puanlamalarda merkeze yönelme etkisinin bulunup bulunmadığına dair nihai karar verilmeden önce, her bir puanlayıcıya ait kategori istatistikleri tablosu incelenmelidir. Başarı testindeki ilk altı soruda, araştırmaya dâhil edilen yedi puanlayıcının her biri için elde edilen kategori istatistikleri, Tablo 4.6'da sunulmuştur.

Tablo 4.6. Standart rubrikler kullanılarak puanlanan ilk altı soruda her bir puanlayıcı için hesaplanan kategori istatistikleri

Puanlayıcı	Kategori İstatistikleri			Kalite Kontrol		Uyguluk Dışı İstatistiği
	Kategoriler	Frekans	Yüzde	Ortalama Ölçüm	Beklenen Ölçüm	
P1	0	222	36	-.98	-1.08	1.4
	1	140	22	-.62	-.52	1.1
	2	146	23	.21	.10	1.2
	3	116	19	.60	.74	1.2
P2	0	144	23	-1.05	-1.13	1.0
	1	222	36	-.75	-.52	.7
	2	146	23	.29	.16	1.1
	3	112	18	.67	.81	1.3
P3	0	171	27	-.79	-.90	1.3
	1	194	31	-.49	-.34	.9
	2	95	15	.25	.27	.9
	3	164	26	.96	.91	.9
P4	0	321	51	-1.42	-1.43	1.1
	1	89	14	-.70	-.83	.8
	2	150	24	-.06	-.17	1.2
	3	64	10	.17	.41	1.4
P5	0	253	41	-1.19	-1.20	1.1
	1	124	20	.81	-.62	1.0
	2	124	20	-.08	.01	1.5
	3	123	20	.63	.60	.09
P6	0	342	55	-1.61	-.1.64	1.1
	1	69	11	-1.11	-1.01	1.1
	2	170	27	-.24	-.34	1.6
	3	43	7	-.05	.23	1.1
P7	0	243	39	-1.47	-1.50	1.1
	1	157	25	-.83	-.88	.8
	2	187	30	-.03	-.18	1.0
	3	37	6	.41	.47	.8

Tablo 4.6'daki bulgular, puanlama ölçeğinin orta kategorilerini yoğun bir biçimde kullanılan puanlayıcı olmadığını göstermektedir. Tablodaki uygunluk dışı istatistiklerinin .5 ile 2.0 kabul edilebilir aralığı içinde kalması, puanlamalarda merkeze yönelme etkisinin bulunmadığını yansıtan bir başka kanıttır. Buna göre, dörtlü derecelendirmeye sahip standart rubrikler kullanılarak puanlanan ilk altı soruda merkeze yönelme etkisinin bulunmadığı söylenebilir. Yedi ve sekiz numaralı sorularda her bir puanlayıcı için elde edilen kategori istatistikleri ise Tablo 4.7'de gösterilmiştir.

Tablo 4.7. Standart rubrikler kullanılarak puanlanan yedi ve sekiz numaralı sorularda her bir puanlayıcı için hesaplanan kategori istatistikleri

Puanlayıcı	Kategori İstatistikleri			Kalite Kontrol		Uyguluk Dışı İstatistiği
	Kategoriler	Frekans	Yüzde	Ortalama Ölçüm	Beklenen Ölçüm	
P1	0	19	9	-.77	-.68	.8
	1	28	13	-.31	-.22	.6
	2	53	25	-.00	.12	1.0
	3	76	37	.48	.40	.7
	4	32	15	.62	.65	1.1
P2	0	17	8	-.86	-.84	2.4
	1	34	16	-.11	-.27	1.7
	2	136	65	.28	.17	.9
	3	17	8	.62	.50	.5
	4	4	2	1.12	.77	.3
P3	0	21	10	-.57	-.50	.9
	1	25	12	.02	-.05	1.0
	2	47	23	.27	.30	1.0
	3	62	30	.62	.57	.6
	4	53	25	.88	.83	.9
P4	0	27	13	-1.25	-1.02	.6
	1	32	15	-.66	-.58	.7
	2	87	42	-.28	-.25	1.0
	3	49	24	.15	.02	.9
	4	13	6	.02*	.27	1.5
P5	0	27	13	-.97	-.95	1.7
	1	27	13	-.46	-.43	1.1
	2	117	56	.08	.04	.9
	3	31	15	.43	.27	.6
	4	6	3	.72	.53	.5
P6	0	27	13	-1.31	-1.30	.9
	1	27	13	-.98	-.83	.7
	2	116	56	-.49	-.48	1.1
	3	34	16	-.10	-.20	1.0
	4	4	2	-.12*	.06	1.2
P7	0	23	11	-1.18	-.87	.3
	1	22	11	-.84	-.43	.2
	2	132	63	-.14	-.10	1.2
	3	31	15	.17	.18	1.6
	4	-	-	-	-	-

Tablo 4.7'ye bakıldığında, bir, üç ve dört numaralı puanlayıcıların puanlama ölçeğinin kategorilerini dengeli bir biçimde kullandığı görülmektedir. Bu puanlayıcılara ait uygunluk dışı istatistikleri de kabul edilebilir sınırlar içerisinde yer almaktadır. Buna göre; bir, üç ve dört numaralı puanlayıcılar tarafından yapılan puanlamalarda merkeze yönelme etkisinin bulunmadığı söylenebilir. Tablo 4.7'ye göre, iki, beş, altı ve yedi numaralı puanlayıcıların puanlama ölçeğinin orta kategorilerini diğer kategorilere kıyasla daha yoğun bir biçimde kullandıkları belirlenmiştir. Puanlama ölçeği kategorilerinin kullanım örüntüsü; iki, beş, altı ve yedi numaralı puanlayıcılar tarafından yapılan puanlamalarda merkeze yönelme

etkisinin bulunduğuna yönelik bir kanaat oluşturmaktadır. Ancak bu konuda; nihai bir karar verilmeden önce uygunluk dışı istatistiklerinin incelenmesi yerinde olacaktır. Uygunluk dışı istatistikleri, beş ve altı numaralı puanlayıcılarda kabul edilebilir sınırlar içerisinde yer alırken; iki ve yedi numaralı puanlayıcılar için kabul edilebilir aralığın dışında kalmıştır. Bu noktadan hareketle, beş ve altı numaralı puanlayıcılar tarafından yapılan puanlamalarda merkeze yönelme etkisi bulunsa da; söz konusu etkinin özellikle iki ve yedi numaralı puanlayıcılar tarafından yapılan puanlamalarda görüldüğü söylenebilir.

4.1.2.3. Halo Etkisi

Halo etkisini belirlemeye yönelik işlemler yürütülürken, puanlayıcı katılığı/cömertliği ile merkeze yönelme etkisinde olduğu gibi, ilk olarak grup düzeyindeki incelemelere yer verilmiştir. Ardından bireysel düzeydeki istatistiksel göstergeler incelenmiştir. Madde yüzeyi için hesaplanan ayırma oranı, ayırma indeksi, güvenirlik indeksi ve Ki Kare değeri, puanlamalarda halo etkisinin bulunup bulunmadığını belirlemek amacıyla kullanılabilecek grup düzeyindeki istatistiksel göstergelerdir. Madde yüzeyine ilişkin ayırma oranı 16.50, ayırma indeksi 22.3 [(4×16.50+1)/3] ve güvenirlik indeksi 1.00 olarak bulunmuştur. Ayırma oranı, ayırma indeksi ve güvenirlik indeksinin yüksek olması maddelerin güçlük düzeyleri açısından farklılık gösterdiğini ortaya koymakta ve puanlamalarda halo etkisinin bulunmadığına işaret etmektedir. Maddelerin güçlük düzeyleri arasında görülen farkın anlamlı olup olmadığına karar vermek için ise Ki Kare değeri incelenmiştir. Ki Kare değerinin istatistiksel olarak anlamlı bulunması [χ^2=1822.22, *sd*=7 *p*<.01], maddelerin güçlük düzeyleri arasındaki farkın manidar olduğunu göstermiştir. Maddelerin güçlük düzeyleri arasında tespit edilen anlamlı fark, puanlamalarda halo etkisinin bulunmadığı şeklinde yorumlanabilir.

Puanlamalarda halo etkisinin bulunup bulunmadığı konusunda karar vermek için yapılabilecek bir diğer işlem; madde yüzeyinin bileşenleri güçlük düzeyleri açısından birbirine eşitlendikten sonra, çok yüzeyli Rasch analizinin tekrarlanması ve puanlayıcılara ait uygunluk istatistiklerinin incelenmesidir. (Linacre, 2014). Maddelerin güçlük düzeyleri açısından eşitlendiği modelde, veri ile mükemmel uyum gösteren puanlayıcıların ölçülen özelliğin farklı yönlerini birbirinden ayırt edemediği kabul edilmektedir. Bu kapsamda, matematik başarı testindeki sekiz soru güçlük düzeyi açısından birbirine eşitlendikten sonra çok yüzeyli Rasch analizi

yinelenmiştir. Yinelenen analizde; puanlayıcılara ait uygunluk içi istatistiklerinin .76 ile 1.17 arasında sıralandığı, uygunluk dışı istatistiklerinin .79 ile 1.18 arasında değiştiği ve veri ile mükemmel uyum gösteren puanlayıcı bulunmadığı belirlenmiştir. Bu bulgu, standart rubriklere göre yapılan puanlamalara halo etkisinin karışmadığını ve öğrencilerin farklı maddelere verdikleri yanıtların birbirinden bağımsız olarak puanlanabildiğini göstermektedir.

Puanlamalarda halo etkisinin bulunup bulunmadığının bireysel düzeydeki göstergesi puanlayıcılara ait uygunluk içi ve uygunluk dışı istatistiklerdir. Uygunluk istatistiklerine ilişkin hangi değerlerin halo etkisine işaret ettiği maddelerin güçlük düzeyleri arasındaki farklılaşmaya göre değişiklik göstermektedir. Maddelerin güçlük düzeyleri arasındaki farklılaşma az olduğunda, 1'den önemli ölçüde düşük olan uygunluk istatistikleri puanlamalara halo etkisi karıştığını yansıtmaktadır. Diğer taraftan; maddelerin güçlük düzeyleri arasındaki değişkenlik fazla olduğunda, uygunluk istatistiklerinin 1'den önemli ölçüde büyük olması puanlamalara halo etkisinin karıştığı anlamına gelmektedir (Myford ve Wolfe, 2004). Standart rubriklere göre yapılan puanlamalarda maddelerin güçlük düzeyleri arasında 2.01 logitlik bir değişim gözlenmiştir. Bu değişim, güçlük düzeyleri açısından maddeler arasındaki farklılaşmanın fazla olduğunu göstermektedir. Dolayısıyla, puanlayıcılara ait uygunluk istatistiklerinin 1'den önemli ölçüde yüksek olması halo etkisine işaret eden bir gösterge olarak kabul edilmiştir. Puanlayıcı yüzeyi için hesaplanan uygunluk içi istatistikleri .76 ile 1.16 arasında değişirken; uygunluk dışı istatistikleri .90 ile 1.21 arasında sıralanmaktadır. Uygunluk istatistiklerinin 1'den önemli ölçüde yüksek olmaması, puanlamalarda halo etkisinin bulunmadığı anlamına gelmektedir. Özetle; hem grup düzeyindeki hem de bireysel düzeydeki istatistiksel göstergeler standart rubrik kullanılarak yapılan puanlamalarda halo etkisinin bulunmadığını ortaya koymuştur.

4.1.2.4. Tutarsızlık Etkisi

Standart rubriklere göre yapılan puanlamalarda tutarsızlık etkisinin bulunup bulunmadığını belirlemeye yönelik işlemlere grup düzeyindeki göstergelerin incelenmesi ile başlanmıştır. Bu kapsamda ilk olarak birey yüzeyinin bileşenleri yetenek düzeyleri açısından birbirine eşitlenerek çok yüzeyli Rasch analizi yinelenmiş ve puanlayıcılara ait uygunluk istatistikleri incelenmiştir. Yinelenen analizde, puanlayıcılar için hesaplanan uygunluk içi istatistiklerinin .81 ile 1.33

arasında değiştiği, uygunluk dışı istatistiklerinin .82 ile 1.42 arasında sıralandığı ve veri ile mükemmel uyum gösteren puanlayıcı bulunmadığı saptanmıştır. Birey yüzeyinin bileşenlerinin yetenek düzeyi açısından eşitlendiği modelde, veri ile mükemmel uyum gösteren puanlayıcıların tutarsız puanlamalar yaptığı ifade edilmektedir (Linacre 2003'ten akt. Myford ve Wolfe, 2003). Dolayısıyla, araştırmaya dâhil edilen puanlayıcılar arasında standart rubrikleri tutarsız bir biçimde kullanan puanlayıcı olmadığı söylenebilir.

Standart rubriklere göre yapılan puanlamalarda tutarsızlık etkisinin bulunup bulunmadığını belirlemek için incelenen grup düzeyindeki diğer göstergeler; birey yüzeyine ilişkin ayırma oranı, ayırma indeksi, güvenirlik indeksi ve Ki Kare değeridir. Birey yüzeyi için elde edilen ölçüm raporlarına bakıldığında; ayırma oranı 3.98, ayırma indeksi 5.64 [(4×3.98+1)/3] ve güvenirlik indeksi .94 olarak bulunmuştur. Bu değerlerin yüksek olması öğrencilerin yetenek düzeyleri açısından ayırt edilebildiğini göstermektedir. Aynı şekilde Ki Kare değerinin anlamlı çıkması [χ^2=1458.1, *sd*=103, *p*<.01] yetenek düzeyleri açısından öğrenciler arasında anlamlı fark bulunduğunu ortaya koymaktadır. Dolayısıyla grup düzeyindeki istatistiksel göstergeler, puanlama ölçeğinin kategorilerini tutarsız bir biçimde kullanan puanlayıcı bulunmadığına yönelik birer kanıt olarak yorumlanabilir.

Standart rubriklere göre yapılan puanlamalarda tutarsızlık etkisinin bulunup bulunmadığını tespit etmeye yönelik çalışmaların ikinci aşamasında bireysel düzeydeki incelemelere geçilmiştir. Bu doğrultuda öncelikle, puanlayıcılar için rapor edilen uygunluk istatistiklerine bakılmıştır. Herhangi bir puanlayıcı için hesaplanan uygunluk istatistiklerinin 1'den önemli ölçüde büyük olması, söz konusu puanlayıcının puanlama ölçeğinin kategorilerini kullanırken tutarsız davrandığına yönelik sinyaller olarak kabul edilmektedir. Standart rubrik kullanılarak yapılan puanlamalarda, puanlayıcılar için hesaplanan uygunluk içi istatistiklerinin .87 ile 1.16 arasında sıralandığı, uygunluk dışı istatistiklerinin ise .97 ile 1.21 arasında değiştiği belirlenmiştir. Uygunluk istatistikleri için rapor edilen bu değerler puanlama ölçeğinin kategorilerini kullanırken, tutarsız davranan puanlayıcı bulunmadığına işaret etmektedir. Tutarsızlık etkisini belirlemeye yönelik bireysel düzeydeki incelemeler kapsamında yapılabilecek bir diğer işlem, farklı puanlayıcılar tarafından yapılan puanlamalar arasındaki korelasyon (çift serili korelasyon) katsayılarının karşılaştırılmasıdır. Örneğin, *n* puanlayıcının görev aldığı bir araştırmada, birinci puanlayıcının yaptığı puanlamalar ile geriye kalan *n-1* puanlayıcı

tarafından yapılan puanlamalar arasındaki korelasyon K_1, ikinci puanlayıcının yaptığı puanlamalar ile geriye kalan *n-1* puanlayıcı tarafından yapılan puanlamalar arasındaki korelasyon K_2…ve *n*. puanlayıcının yaptığı puanlamalar ile geriye kalan *n-1* puanlayıcı tarafından yapılan puanlamalar arasındaki korelasyon K_n olsun. Puanlamalarda tutarsızlık etkisinin bulunup bulunmadığını belirlemek için bu *n* tane korelasyon katsayısı karşılaştırılır. Bu sayede, diğerlerine göre anlamlı derecede daha düşük olan korelasyon katsayıları belirlenir. Belirlenen bu korelasyon katsayılarının karşılık geldiği puanlayıcıların puanlama ölçeğini diğer puanlayıcılara göre daha farklı yorumladığı ve tutarsız puanlamalar yaptığı kabul edilmektedir (Myford ve Wolfe, 2004). Standart rubriklere göre yapılan puanlamalar için hesaplanan korelasyon katsayılarının .43 ile .48 aralığında bulunduğu belirlenmiştir. Hesaplanan yedi farklı korelasyon katsayısının hiçbiri diğerlerine göre anlamlı derecede daha düşük olmadığından, tutarsız puanlamalar yapan puanlayıcı olmadığı ifade edilebilir. Hung, Chen ve Chen'e göre (2012), hesaplanan korelasyon katsayılarının orta düzeyde olması da puanlama ölçeğinin kategorilerini diğer puanlayıcılara göre daha farklı yorumlayan puanlayıcı olmadığı anlamına gelmektedir. Standart rubrik kullanılarak yapılan puanlamalarda tutarsızlık etkisi bulunup bulunmadığını belirlemek amacıyla gerek grup düzeyinde gerekse de bireysel düzeyde yapılan incelemeler, puanlama ölçeğinin kategorilerini tutarsız bir biçimde kullanan puanlayıcı bulunmadığını ortaya koymaktadır.

4.1.2.5. Yanlılık Etkisi

Puanlayıcıların standart rubrik kullanılarak yaptıkları puanlamalarda, öğrencilerin bir kısmına diğerlerine göre daha katı ya da daha cömert davranıp davranmadığını belirlemek için puanlayıcı ile birey etkileşimi (*puanlayıcı×birey*) incelenmiştir. Puanlayıcıların testteki maddelerin tamamını aynı katılık ya da cömertlikte puanlayıp puanlamadığını tespit etmek için ise puanlayıcı ve madde yüzeyleri arasındaki etkileşime (*puanlayıcı×madde*) bakılmıştır. Puanlayıcıların, öğrencilerin bir kısmının bazı maddelerdeki performansına beklenenden daha düşük ya da yüksek puanlar verip vermediğini saptamak için ise puanlayıcı, birey ve madde yüzeyleri arasındaki etkileşim (*puanlayıcı×birey×madde*) incelenmiştir. Araştırmada puanlayıcı etkileri ile ilgilenildiğinden, bazı öğrencilerin testteki maddelerin bir kısmında beklenenden daha yüksek ya da daha düşük bir performans gösterip

göstermediği kapsam dışı tutulmuştur. Dolayısıyla, birey ve madde yüzeyleri arasındaki etkileşim (*birey*×*madde*) araştırmaya dâhil edilmemiştir.

Standart rubriklere göre yapılan puanlamalarda, puanlayıcıların öğrencilerden bazılarını diğerlerine göre daha katı ya da daha cömert değerlendirip değerlendirmediğini belirlemek için *puanlayıcı*×*birey* etkileşimleri incelenmiştir. *Puanlayıcı*×*birey* etkileşimlerinin anlamlılığının belirlenmesinde elde edilen *t* değerlerinden yararlanılmaktadır. ±2 aralığının dışında kalan *t* değerinin anlamlı etkileşime işaret ettiği kabul edilmektedir (Linacre, 2014). Araştırmada, 104 öğrencinin açık uçlu matematik sorularına verdikleri yanıtlar yedi puanlayıcı tarafından puanlamıştır. Dolayısıyla puanlayıcı ve birey yüzeyleri arasında 728 (7×104) olası etkileşim bulunmaktadır. Yanlılık analizi sonucunda hesaplanan *t* testi değerleri bu 728 etkileşimden üçünün istatistiksel olarak anlamlı olduğunu göstermiştir. Anlamlı olduğu tespit edilen *puanlayıcı*×*birey* etkileşimleri Tablo 4.8'de sunulmuştur. Puanlayıcı ve birey yüzeyleri arasındaki 728 etkileşimin tümüne ilişkin analiz çıktılarının Tablo 4.8'de gösterilmesi mümkün olmadığından, tabloda yalnızca iki yüzey arasındaki anlamlı etkileşimlere yer verilmiştir.

Tablo 4.8. Standart rubrik kullanılarak yapılan puanlamalarda anlamlı olduğu tespit edilen puanlayıcı×birey etkileşimleri

Puanlayıcı	Birey	Gözlenen Puan	Beklenen Puan	Bias (Logit)	Standart Hata	*t*
P3	79	25	19.27	2.11	1.02	2.08
P6	50	10	4.87	1.02	.42	2.40
P3	53	19	13.78	.89	.44	2.04

Ki Kare=274.8, *sd*= 728, *p*<.05

Tablo 4.8 incelendiğinde; üç numaralı puanlayıcının 79 ve 53 numaralı öğrencileri puanlarken, altı numaralı puanlayıcının ise 50 numaralı öğrenciyi puanlarken beklenenden yüksek puanlamalar yaptığı görülmektedir. Buna göre, puanlayıcı ve birey yüzeyleri arasındaki olası 728 etkileşimden yalnızca üçü (%0.4) istatistiksel olarak anlamlı bulunmuştur. Tablo 4.8'e göre, "*puanlayıcı*×*birey*" yanlılığına ilişkin Ki Kare değeri istatistiksel açıdan anlamlı değildir [χ^2=274.8, *sd*=728, *p*>.05]. Dolayısıyla, standart rubriklere göre yapılan puanlamalarda "*puanlayıcı*×*birey*" şeklinde bir yanlılığın bulunmadığı söylenebilir. Diğer bir ifadeyle, puanlayıcılar standart rubriklere göre puanlama yaparken; öğrencilerin bir kısmına diğerlerine göre daha katı ya da daha cömert davranmamış; tüm öğrencileri aynı katılık/cömertlikte değerlendirmiştir.

Puanlayıcı×*madde* etkileşimine ilişkin elde edilen bulgular Tablo 4.9 ve Tablo 4.10'da sunulmuştur. Puanlayıcı ve madde yüzeyleri arasında 56 [7 (puanlayıcı) × 8 (madde)] olası etkileşim söz konusudur. Puanlayıcı ve madde yüzeyleri arasındaki etkileşimlerin sayısı, puanlayıcı ve birey yüzeyleri arasındaki etkileşim sayısına göre çok daha azdır. Bundan dolayı, puanlayıcı ve madde yüzeyleri arasındaki olası 56 etkileşimin her biri için hesaplanan *t* değerleri Tablo 4.9'da sunulabilmiştir.

Tablo 4.9. Standart rubrik kullanılarak yapılan puanlamalarda puanlayıcı ve madde yüzeyleri arasındaki etkileşime ilişkin *t* değerleri

		Puanlayıcılar						
		P1	P2	P3	P4	P5	P6	P7
Maddeler	M1	-2.72*	1.99	-2.08*	0.13	0.09	0.42	2.8*
	M2	-3.24*	4.38*	-0.36	-1.94	1.35	0.35	-0.55
	M3	-1.27	0.69	-1.73	1.94	1.31	-1.16	0.41
	M4	-1.42	0.74	-3.07*	-1.62	2.15*	1.14	2.34*
	M5	0.09	-2.78*	1.72	0.34	2.6*	-0.66	-1.13
	M6	4.59*	3.09*	2.68*	-2.3*	-3.51*	-3.32*	-2.81*
	M7	2.55*	-3.72*	2.66*	0.93	-3.32*	1.78	-0.86
	M8	1.31	-4.17*	0.62	2.26*	-0.75	0.67	0.03

*$|t| \geq 2$

Tablo 4.9'da görüldüğü gibi, puanlayıcı ve madde yüzeyleri arasındaki 56 etkileşimden 23'ü (%41) istatistiksel olarak anlamlıdır. *Puanlayıcı*×*madde* yanlılığına ilişkin Ki Kare değerinin anlamlı olması [χ^2=263.50, *sd*=56, p<.05], standart rubriklere göre yapılan puanlamalarda *puanlayıcı*×*madde* yanlılığının bulunduğunu göstermektedir. Buna göre, puanlayıcıların bazı maddeleri puanlarken beklenenden daha katı ya da daha cömert davrandığı söylenebilir. Puanlayıcı ve madde yüzeyleri arasındaki anlamlı etkileşimlere ilişkin ayrıntılı bulgular Tablo 4.10'da gösterilmiştir.

Tablo 4.10. Standart rubrik kullanılarak yapılan puanlamalarda anlamlı olduğu tespit edilen puanlayıcı×madde etkileşimleri

Puanlayıcı	Madde	Gözlenen Puan	Beklenen Puan	Bias (Logit)	Standart Hata	*t*
P2	2	240	201.77	.59	.13	4.38
P1	6	122	82.95	.50	.11	4.59
P7	1	70	50.17	.35	.13	2.80
P2	6	103	77.41	.35	.11	3.09
P5	5	235	213.29	.34	.13	2.60
P3	7	267	243.63	.31	.12	2.66
P3	6	122	98.34	.29	.11	2.68
P1	7	250	227.39	.29	.11	2.55
P7	4	106	86.05	.26	.11	2.34
P4	8	209	188.60	.25	.11	2.26
P5	4	118	99.03	.24	.11	2.15
P3	1	72	89.93	-.26	.12	-2.08
P2	5	203	225.27	-.32	.12	-2.78
P3	4	109	137.51	-.34	.11	-3.07
P1	2	180	207.63	-.36	.11	-3.24
P5	7	177	206.74	-.37	.11	-3.32
P4	6	36	52.59	-.38	.16	-2.30
P1	1	53	75.32	-.38	.14	-2.72
P2	7	188	221.06	-.41	.11	-3.72
P2	5	185	222.05	-.46	.11	-4.17
P7	6	35	55.92	-.47	.17	-2.81
P5	6	38	65.86	-.56	.16	-3.51
P6	6	23	46.23	-.67	.20	-3.52

Ki Kare=263.50, *sd*=56, $p<.05$

Tablo 4.10'a göre, bir numaralı puanlayıcı altı ve yedi numaralı maddelerde beklenenden daha yüksek; bir ve iki numaralı maddelerde ise beklenenden daha düşük puanlamalar yapmıştır. İki numaralı puanlayıcının, iki numaralı maddede beklenenden daha yüksek; beş, altı, yedi ve sekiz numaralı maddelerde beklenenden daha düşük puanlamalar yaptığı tespit edilmiştir. Üç numaralı puanlayıcı bir ve dört numaralı maddelerde beklenenden düşük; altı ve yedi numaralı maddelerde beklenenden daha yüksek puanlamalar yapmıştır. Dört numaralı puanlayıcının altı numaralı maddedeki puanlamaları beklenenden daha düşük; sekiz numaralı maddedeki puanlamaları ise beklenenden daha yüksek bulunmuştur. Beş numaralı puanlayıcının dört ve beş numaralı maddelerdeki puanlamalarının beklenenden daha yüksek; altı ve yedi numaralı maddelerdeki puanlamalarının ise beklenenden daha düşük olduğu belirlenmiştir. Altı numaralı puanlayıcının altı numaralı maddede beklenenden daha düşük puanlamalar yaptığı saptanmıştır. Son olarak, yedi numaralı puanlayıcının bir ve dört numaralı maddelerde beklenenden daha yüksek; altı numaralı maddede ise beklenenden daha düşük puanlamalar yaptığı tespit edilmiştir.

Puanlayıcı×*birey*×*madde* etkileşimine ilişkin elde edilen bulgular Tablo 4.11'de sunulmuştur. Bu üç yüzey arasında 5824 [7 (puanlayıcı) × 104 (birey) × 8

(madde)] olası etkileşim bulunmaktadır. Yanlılık analizi sonucunda elde edilen *t* değerleri, 5824 etkileşimden, yalnızca 27'sinin istatistiksel olarak anlamlı olduğunu göstermiştir. Puanlayıcı, birey ve madde yüzeyleri arasındaki 5824 etkileşimin tümüne ilişkin analiz çıktılarının Tablo 4.11'de gösterilmesi mümkün olmadığından, tabloda yalnızca üç yüzey arasındaki anlamlı etkileşimlere yer verilmiştir. Tablo 4.11'deki Ki Kare değerinin anlamlı olmaması [χ^2=2893, *sd*=5824, *p*>.05], standart rubriklere göre yapılan puanlamalarda, *puanlayıcı*×*birey*×*madde* yanlılığının bulunmadığını yansıtmaktadır.

Tablo 4.11. Standart rubrik kullanılarak yapılan puanlamalarda anlamlı olduğu tespit edilen puanlayıcı×birey×madde etkileşimleri

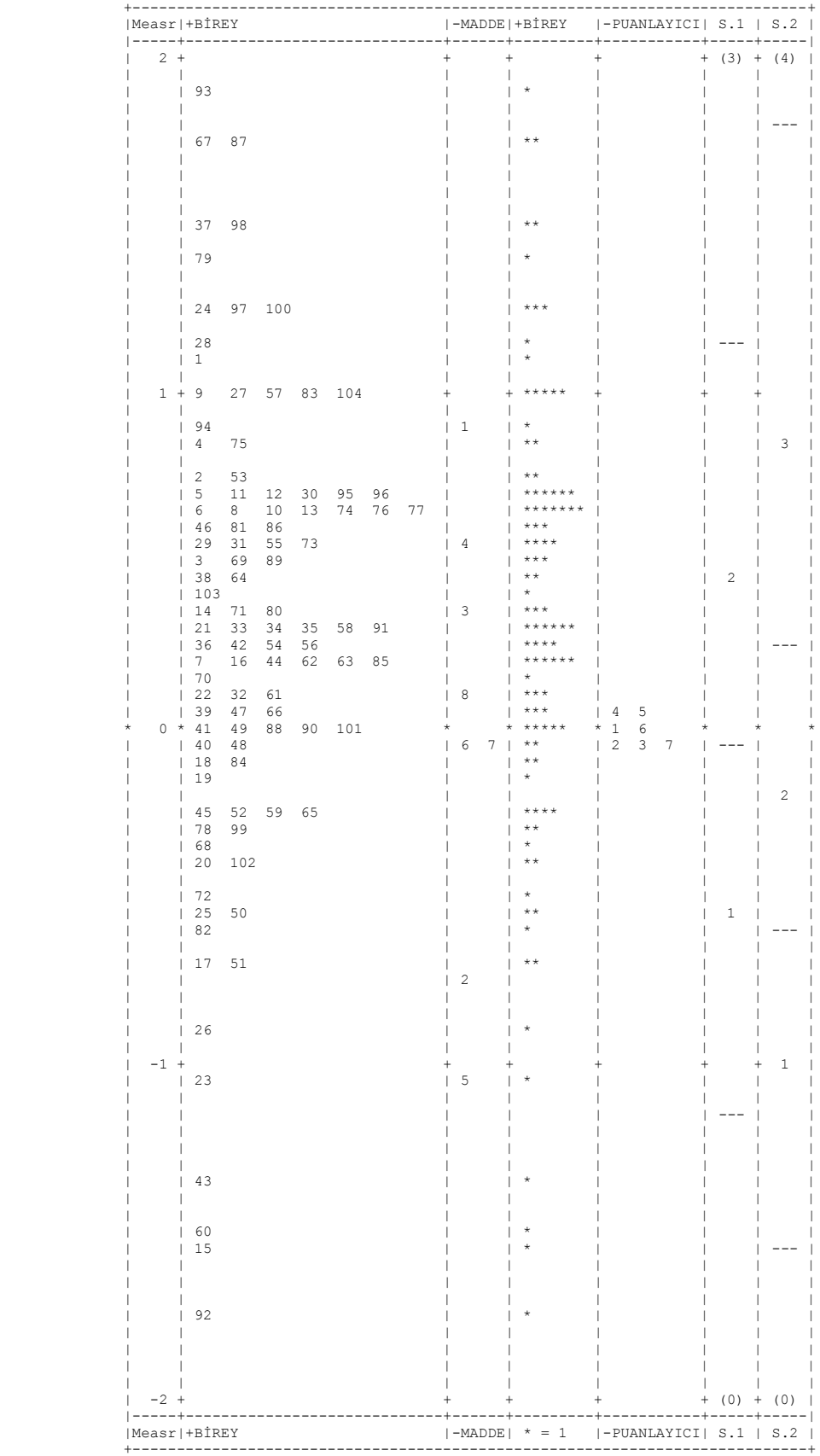
Puanlayıcı	Birey	Madde	Gözlenen Puan	Beklenen Puan	Bias (Logit)	Standart Hata	*t*
P6	92	4	2	.06	3.82	1.12	3.39
P7	92	4	2	.07	3.62	1.12	3.22
P6	23	1	2	.07	3.58	1.12	3.19
P4	23	1	2	.09	3.45	1.12	3.07
P5	92	4	2	.09	3.45	1.12	3.07
P7	23	1	2	.09	3.39	1.12	3.01
P2	92	4	2	.10	3.27	1.12	2.91
P1	92	4	2	.11	3.19	1.12	2.83
P6	17	3	2	.12	3.15	1.12	2.80
P4	17	3	2	.13	3.02	1.12	2.69
P7	17	3	2	.14	2.96	1.12	2.63
P2	15	4	2	.20	2.65	1.12	2.36
P6	39	1	2	.22	2.57	1.12	2.29
P1	17	3	2	.23	2.53	1.12	2.25
P6	99	6	2	.25	2.46	1.12	2.19
P4	39	1	2	.25	2.44	1.12	2.17
P4	18	1	2	.27	2.39	1.12	2.12
P7	80	1	2	.27	2.38	1.12	2.11
P5	32	1	2	.27	2.37	1.12	2.10
P6	66	1	2	.28	2.34	1.12	2.08
P4	99	6	2	.28	2.33	1.12	2.07
P7	99	6	2	.30	2.27	1.12	2.01
P3	1	5	1	2.60	-2.15	1.07	-2.01
P1	79	2	1	2.61	-2.20	1.07	-2.05
P2	98	2	0	2.64	-3.00	1.39	-2.15
P1	98	2	0	2.66	-3.53	1.75	-2.02
P3	98	2	0	2.72	-3.66	1.68	-2.18

Ki Kare= 2893, *sd*= 5824, *p*<.05

4.1.3. Üçüncü Alt Probleme İlişkin Bulgular

Araştırmanın üçüncü alt problemi "SOLO taksonomisine dayalı rubriklere göre yapılan puanlamalarda; puanlayıcı, birey (öğrenci) ve madde yüzeyleri için hesaplanan güvenirlik değerleri ile uygunluk istatistikleri nasıldır?" şeklinde ifade edilmiştir. Bu alt problem için, öğrencilerin açık uçlu matematik sorularına verdikleri yanıtlar yedi puanlayıcı tarafından puanlanmış, puanlama sonucunda elde edilen

veriler çok yüzeyli Rasch modeline göre analiz edilmiştir. Analiz çıktılarında ilk olarak verilerin program tarafından kaç defa okunduğunu gösteren iterasyon raporu yer almıştır. İterasyon raporu incelendiğinde, verilerin program tarafından 21 defa okunduğu tespit edilmiştir. İterasyon sayısının az olması veriden iyi kestirimde bulunmanın kolay olduğunu göstermektedir. İterasyon raporunun ardından değişken haritası incelenmiştir. Değişken haritası Şekil 4.4'te gösterilmiştir. Şekil 4.4 incelendiğinde, ilk sütunda logit ölçeğinin bulunduğu görülmektedir. Değişken haritasının ikinci sütununda bireyler yetenek düzeyleri açısından sıralanmıştır. Logit ölçeğinin alt ucundan üst ucuna doğru gidildikçe öğrencilerin yetenek düzeyi artmaktadır. Buna göre, en yüksek yetenek düzeyindeki öğrencinin 93 numaralı öğrenci (1.91 logit), en düşük yetenek düzeyindeki öğrencinin ise, 92 numaralı öğrenci (-1.74 logit) olduğu görülmektedir. Değişken haritasının üçüncü sütununda maddeler bulunmaktadır. Logit ölçeğinin negatif ucundaki maddelerin daha kolay, pozitif ucundaki maddelerin ise daha zor olduğu bilinmektedir. Dolayısıyla, testteki en zor sorunun bir numaralı madde (.88 logit); en kolay sorunun ise beş numaralı madde (-1.04 logit) olduğu ortaya çıkmaktadır. Ayrıca, altı ve yedi numaralı maddelerin yaklaşık olarak aynı güçlükte olduğu görülmektedir. Değişken haritasının beşinci sütununda ise puanlayıcılar yer almıştır. Katı puanlamalar yapan puanlayıcılar logit ölçeğinin üst kısmında yer alırken; cömert puanlamalar yapan puanlayıcılar logit ölçeğinin alt kısmında yer almaktadır. Şekil 4.4'e göre, puanlayıcıların genel olarak logit ölçeğinin sıfır noktasında bulunduğu veya bu noktaya oldukça yakın olduğu saptanmıştır. Birey ve madde yüzeyleri için logit ölçeğinin alt ve üst ucu boyunca uzanan ölçümlerin elde edilmesi, bireylerin yetenek düzeyleri açısından ayırt edilebildiğine ve maddelerin güçlük düzeyleri açısından farklılık gösterdiğine işaret etmektedir. Puanlayıcıların logit ölçeğinin orta noktasında kümelenmesi ise, puanlama işlemindeki katılık ve cömertlikleri açısından puanlayıcılar arasında önemli bir farkın bulunmadığını düşündürmektedir. Ancak; yetenek düzeyleri açısından bireyler arasında, güçlük düzeyleri açısından maddeler arasında ve katılık/cömertlikleri açısından puanlayıcılar arasında anlamlı bir fark olup olmadığını ortaya koyabilmek için her bir yüzeye ilişkin ölçüm raporlarının incelenmesi gerekmektedir.



Şekil 4.4. SOLO taksonomisine dayalı rubriklere göre yapılan puanlamaların çok yüzeyli Rasch modeline göre analiz edilmesiyle elde edilen değişken haritası

Puanlayıcı yüzeyine ilişkin ölçüm raporları Tablo 4.12'de sunulmuştur. Tablo 4.12'ye göre, puanlayıcılara ilişkin logit ölçüleri .06 ile -.06 arasında değişmekte olup, puanlayıcıların katılık ve cömertliklerine ilişkin aralık .12 logittir [.06-(-.06)]. Bu aralığın küçük olması, katılık ve cömertlikleri açısından puanlayıcılar arasında önemli bir fark bulunmadığını düşündürmektedir. Puanlayıcılar için rapor edilen uygunluk içi ve uygunluk dışı istatistiklerinin ortalamasına bakıldığında, .99 ve 1.12 gibi 1'e oldukça yakın değerler olduğu belirlenmiştir. Bu değerler verinin model ile uyumlu olduğunu göstermektedir. Ayrıca, uygunluk içi ve uygunluk dışı istatistiklerinin puanlayıcıların hiçbirinde .5 ile 2.00 kabul edilebilir aralığının (Myford ve Wolfe, 2003) dışında kalmadığı saptanmıştır. Bu bulgu, model ile veri uyumunu olumsuz etkileyen puanlayıcı bulunmadığı anlamına gelmektedir.

Tablo 4.12'ye göre, puanlayıcı yüzeyine ilişkin ayırma oranı .69 ve güvenirlik indeksi .32 olarak belirlenmiştir. Birey ve madde yüzeylerinde güvenirlik indeksinin yüksek olması istenilen bir durum iken; puanlayıcı yüzeyi söz konusu olduğunda güvenirlik indeksinin düşük olması istenmektedir. Çünkü puanlayıcı yüzeyi için hesaplanan güvenirlik indeksi, puanlayıcılar arasındaki farkın güvenirliğidir. Bu katsayının yüksek olması, katılık ve cömertlikleri yönüyle puanlayıcılar arasında fark bulunduğunu göstermektedir. Dolayısıyla, puanlayıcı yüzeyi için hesaplanan ayırma oranı ve güvenirlik indeksinin düşük olması, katılık ve cömertlikleri açısından puanlayıcılar arasında fark olmadığına işaret etmektedir. Bununla birlikte, bu konudaki nihai karar, puanlayıcılar arasındaki farkın istatistiksel olarak anlamlı olup olmadığını yansıtan Ki Kare değeri incelenerek verilmektedir (Linacre, 2014). Ki Kare değerinin istatistiksel açıdan anlamlı olmaması [χ^2=8.8, *sd*= 6, *p*>.05], katılık ve cömertlikleri açısından puanlayıcılar arasında anlamlı fark bulunmadığına ilişkin bir kanıt sunmaktadır. Puanlayıcılar arası uyum değerleri incelendiğinde, beklenen uyum %37.8 ve gözlenen uyum %82.2 olarak hesaplanmıştır. Puanlayıcılar arasındaki mutlak uyumun yüksek olması, puanlayıcıların aynı öğrenci cevabına büyük ölçüde aynı puanı verdiği anlamına gelmektedir. Puanlayıcılar arası mutlak uyuma ilişkin kabul edilebilir alt sınır %75 olarak belirtilmektedir (Graham, Milanowski ve Miller, 2012). Buna göre, SOLO taksonomisi kullanılarak yapılan puanlamalarda puanlayıcılar arası güvenirliğin yeterli düzeyde olduğu söylenebilir. Ancak mutlak uyum hesaplanırken, puanlayıcılar arasında şansa bağlı olarak oluşabilecek uyum dikkate alınmamaktadır. Şans sonucu ulaşabilecek uyum çıkarıldıktan sonra puanlayıcılar arasında kabul edilebilir uyum olup olmadığının tespiti için Kappa

istatistiği hesaplanmalıdır. Tablo 4.12'ye göre, kappa istatistiği .71 olarak elde edilmiştir. Kappa istatistiği için .75 ve üzerindeki değerler mükemmel uyumu, .60 ile .74 arasındaki değerler ise iyi uyumu yansıtmaktadır (Landis ve Knoch, 1977). SOLO taksonomisine dayalı rubrikler kullanılarak yapılan puanlamalara ait kappa istatistiği bu ölçütler doğrultusunda ele alındığında, puanlayıcılar arasındaki uyumun iyi olduğu söylenebilir.

Tablo 4.12. SOLO taksonomisine dayalı rubriklere göre yapılan puanlamaların çok yüzeyli Rasch modeline göre analiz edilmesiyle puanlayıcı yüzeyi için elde edilen ölçüm raporları

Puanlayıcı	Logit Ölçüsü	Standart Hata	Uygunluk İçi	Uygunluk Dışı
P5	.06	.04	1.02	1.19
P4	.04	.04	1.04	1.16
P1	.02	.04	.98	1.12
P6	.02	.04	1.00	1.15
P3	-.04	.04	1.04	1.12
P7	-.05	.04	.96	1.06
P2	-.06	.04	.91	1.00
Ortalama	.00	.04	.99	1.12
Standart Sapma (Evren)	.04	.00	.04	.06
Standart Sapma (Örneklem)	.05	.00	.05	.07
Model, Evren: RMSE=.04		Standart Sapma=.02	Ayırma Oranı=.51	Güvenirlik=.21
Model, Örneklem: RMSE=.04		Standart Sapma=.03	Ayırma Oranı=.69	Güvenirlik=.32
Model, Tamamı Aynı Ki Kare=8.8 sd=6 p=.18 Model, Rastgele Normal Ki Kare=3.6 sd=5 p=.61				
Puanlayıcılar arası mutlak uyum: %82.2 Puanlayıcılar arası beklenen uyum: %37.8 Puanlayıcılar Arası Güvenirliğe İlişkin Kappa İstatistiği: .71				

Birey yüzeyine ilişkin ölçüm raporları, Tablo 4.13'te gösterilmiştir. Birey yüzeyinde 104 bileşen bulunduğundan, tabloda her bir öğrenciye ilişkin ölçüm raporları ayrı ayrı verilmemiştir. Tabloda, 104 öğrenci için hesaplanan yetenek kestirimleri ve standart hata değerleri ile uygunluk içi ve uygunluk dışı istatistiklerinin ortalaması sunulmuş; ayırma oranı, güvenirlik indeksi ve Ki Kare değerlerine yer verilmiştir.

Tablo 4.13. SOLO taksonomisine dayalı rubriklere göre yapılan puanlamaların çok yüzeyli Rasch modeline göre analiz edilmesiyle birey yüzeyi için elde edilen ölçüm raporları

	Logit Ölçüsü	Standart Hata	Uygunluk İçi	Uygunluk Dışı
Ortalama	.30	.15	1.01	1.12
Standart Sapma (Evren)	.68	.02	.58	.87
Standart Sapma (Örneklem)	.68	.02	.58	.87
Model, Evren: RMSE=.16	Standart Sapma=.66		Ayırma Oranı=4.26	Güvenirlik=.95
Model, Örneklem: RMSE=.16	Standart Sapma=.66		Ayırma Oranı=4.28	Güvenirlik=.95
Model, Tamamı Aynı Ki Kare=1610.4 sd=103 p=.00 Model, Rastgele Normal Ki Kare=96.4 sd=102 p=.64				

Birey yüzeyine ait ölçüm raporlarına göre, öğrencilerin yetenek düzeylerine ilişkin kestirimlerin 1.91 logit ile -1.74 logit arasında değiştiği belirlenmiştir. Bu değerler, öğrencilerin yetenek düzeylerinin 3.65 [1.91-(-1.74)] logit arasında değiştiğini göstermektedir. Öğrencilere ilişkin yetenek kestirimlerinin 3.65 logit gibi geniş bir aralık boyunca uzanması, yetenek düzeyleri farklı olan öğrencilerin birbirinden başarılı bir şekilde ayırt edilebildiğini yansıtmaktadır. Bireylerin yetenek düzeylerinin ortalaması .30 ve standart sapması .68 logit olarak hesaplanmıştır. Uygunluk içi ve uygunluk dışı istatistiklerinin ortalaması sırasıyla 1.01 ve 1.12 şeklindedir. Uygunluk içi ve uygunluk dışı istatistiklerinin ortalaması, bu istatistiklere ilişkin beklenen değer olan 1'e oldukça yakındır. Dolayısıyla, birey yüzeyi için elde edilen uygunluk istatistikleri, model ile veri arasındaki uyumunun iyi olduğuna yönelik bir kanıt olarak yorumlanabilir. Birey yüzeyi için ayırma oranı 4.28 ve güvenirlik indeksi .95 olarak hesaplanmıştır. Bu değerler, öğrencilerin yüksek güvenirlikte birbirinden ayırt edilebildiğine işaret etmektedir. Birey yüzeyi için hesaplanan Ki Kare değerinin anlamlı olması [χ^2=1610.4, *sd*=103, *p*<.01], yetenek düzeyleri açısından öğrenciler arasında anlamlı fark bulunduğunu göstermektedir. Diğer bir deyişle, yetenek düzeyleri farklı olan öğrenciler birbirinden ayırt edilebilmiştir.

Madde yüzeyine ilişkin ölçüm raporları ise Tablo 4.14'te sunulmuştur. Tablo 4.14'e göre, testteki maddelerin güçlük düzeyleri .88 logit ile -1.04 logit arasında sıralanmaktadır. Buna göre, maddelerin güçlük düzeyleri arasında 1.92 [.88-(-1.04)] logitlik bir değişim gözlenmektedir. Maddelerin logit değerlerinin ortalaması .00 ve standart sapması .64 logit olarak bulunmuştur. Maddelere ilişkin uygunluk içi ve

uygunluk dışı istatistiklerinin ortalaması sırasıyla 1.05 ve 1.12 şeklindedir. Bu değerler, uygunluk istatistiklerinin beklenen değeri olan 1'e oldukça yakındır. Dolayısıyla, model ile veri uyumunun sağlandığı belirlenmiştir. Ayrıca testte, Myford ve Wolfe (2003) tarafından uygunluk istatistiklerine ilişkin dikkate alınması önerilen .05 ile 2.0 aralığının dışında kalan madde olmadığı tespit edilmiştir. Testte yer alan iki numaralı maddenin uygunluk dışı istatistiği kabul edilebilir üst sınır olan 2 değerini aşmaktadır. Ancak, uygunluk dışı istatistiklerinin uç değerlere karşı hassas olduğu bilinmekte ve bu nedenle model ile veri arasındaki uyumu değerlendirirken daha çok uygunluk içi istatistiklerinin dikkate alınması önerilmektedir (Bond ve Fox, 2001). Bu noktadan hareketle, testte model ile veri arasındaki uyumu olumsuz yönde etkileyen bir madde bulunmadığı söylenebilir.

Tablo 4.14. SOLO taksonomisine dayalı rubriklere göre yapılan puanlamaların çok yüzeyli Rasch modeline göre analiz edilmesiyle madde yüzeyi için elde edilen ölçüm raporları

Madde	Logit Ölçüsü	Standart Hata	Uygunluk İçi	Uygunluk Dışı
1	.88	.04	1.26	1.23
4	.57	.04	.71	1.01
3	.35	.04	.85	.90
8	.12	.04	1.08	1.10
7	-.05	.04	.62	.60
6	-.07	.04	.78	.82
2	-.76	.05	1.76	2.22
5	-1.04	.06	1.29	1.04
Ortalama	.00	.04	1.05	1.12
Standart Sapma (Evren)	.60	.01	.36	.45
Standart Sapma (Örneklem)	.64	.01	.38	.48
Model, Evren: RMSE=.04	Standart Sapma=.60	Ayırma Oranı=13.93	Güvenirlik=.99	
Model, Örneklem: RMSE=.04	Standart Sapma=.64	Ayırma Oranı=14.90	Güvenirlik=1.00	
Model, Tamamı Aynı Ki Kare=1316.4 sd=7 p=.00				
Model, Rastgele Normal Ki Kare=7.0 sd=6 p=.32				

Tablo 4.14'e göre, madde yüzeyine ilişkin ayırma oranı 14.90 ve güvenirlik indeksi 1.00 olarak elde edilmiştir. Madde yüzeyi için hesaplanan ayırma oranı ve güvenirlik indeksinin yüksek olması, testteki maddelerin güçlük düzeyleri açısından farklılık gösterdiğini yansıtmaktadır. Hesaplanan güvenirlik indeksi, maddelerin güçlük düzeyleri açısından farklılaştığına işaret etse de; bu farkın anlamlı olup olmadığına yönelik karar Ki Kare testi incelenerek verilmektedir. Ki-Kare testinin

istatistiksel açıdan anlamlı çıkması [χ^2=1316.4, *sd*=7, *p*<.01], güçlük düzeyleri açısından maddeler arasında anlamlı fark bulunduğunu göstermektedir.

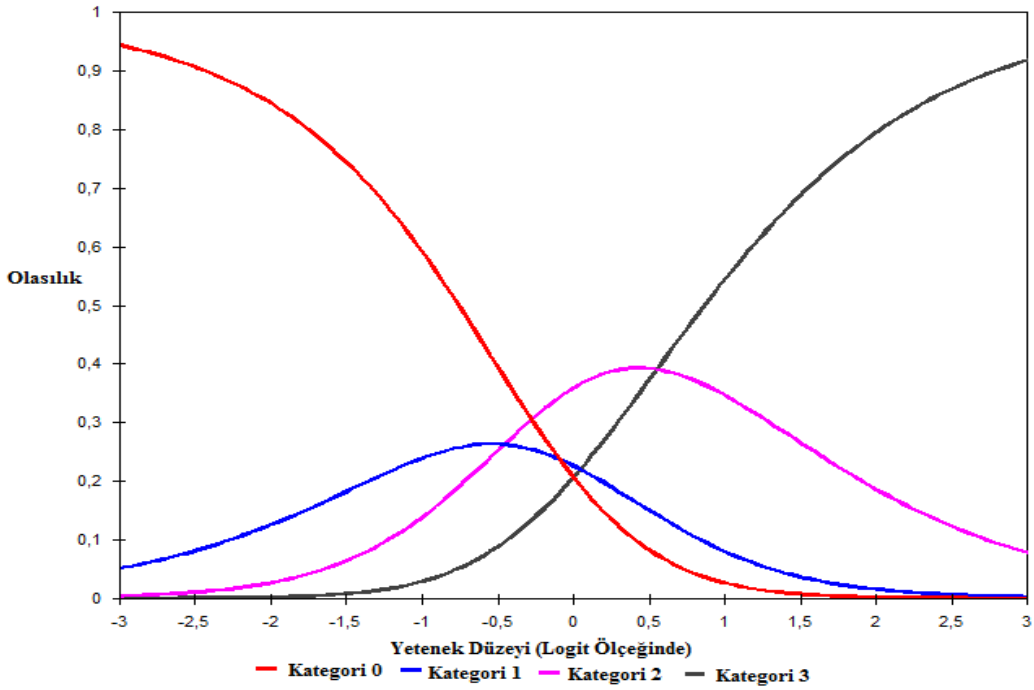
Analiz çıktılarında; iterasyon raporu, değişken haritası, puanlayıcı, birey ve madde yüzeyine ilişkin ölçümlerden sonra kategori istatistikleri yer almaktadır. SOLO taksonomisine dayalı rubrik kullanılarak yapılan puanlamalardan elde edilen kategori istatistikleri Tablo 4.15'te gösterilmiştir. Tablo 4.15'te, başarı testindeki ilk altı sorunun değerlendirilmesinde kullanılan dörtlü derecelendirmeye sahip rubrikler ile yedi ve sekiz numaralı soruların değerlendirilmesinde kullanılan beşli derecelendirmeye sahip rubriklere ilişkin kategori istatistikleri bir arada sunulmuştur.

Tablo 4.15. SOLO taksonomisine dayalı rubriklere göre yapılan puanlamaların çok yüzeyli Rasch modeline göre analiz edilmesiyle elde edilen kategori istatistikleri

Maddeler	Puanlama Ölçeği Kategorileri	Frekans	Yüzde	Yığılmalı Yüzde	Ortalama Ölçüm	Beklenen Ölçüm	Dış Uyum Kareler Ortalaması
1-6 numaralı maddeler	0	918	21	21	-.39	-.63	1.9
	1	695	16	37	-.42*	-.12	.6
	2	1268	29	66	.30	.42	.9
	3	1487	34	100	1.10	1.00	1.00
7 ve 8 numaralı maddeler	0	149	10	10	-.67	-.62	.8
	1	117	8	18	-.24	-.16	.8
	2	426	29	47	.06	.17	.8
	3	508	35	82	.59	.45	.7
	4	256	18	100	.73	.75	1.0

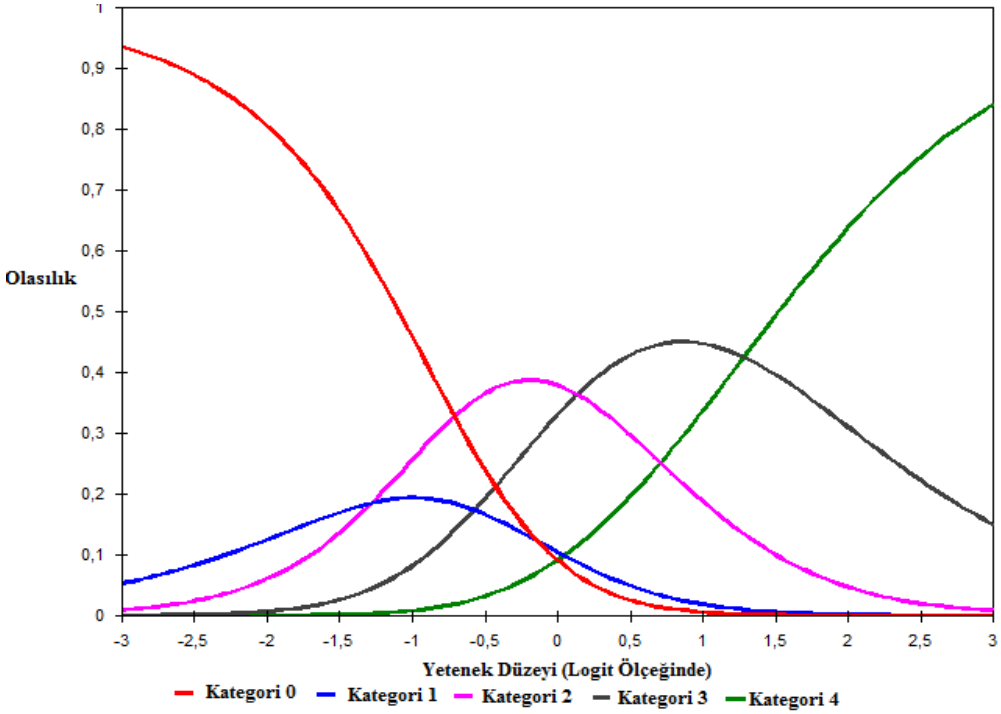
Puanlamada kullanılan rubriklerin etkin bir biçimde çalıştığının söylenebilmesi için karşılanması gereken ilk varsayım, puanlama ölçeğinin her bir kategorisinde en az 10 gözlem bulunmasıdır. Puanlama ölçeğinin kategorilerine karşılık gelen frekans değerleri incelendiğinde, her bir kategorinin en az 10 defa kullanıldığı görülmektedir. Dolayısıyla, puanlama ölçeğinin etkili bir biçimde çalıştığını söyleyebilmek için gerekli olan ilk varsayımın hem dörtlü hem de beşli bir derecelendirmeye sahip rubriklerde karşılandığı söylenebilir. Puanlama ölçeğinin etkin bir biçimde çalıştığına kanaat getirebilmek için karşılanması gereken diğer şartlar; puanlama ölçeğinin kategorileri ile birlikte ortalama ölçümlerin de artması ve uygunluk dışı istatistiklerinin .05 ile 2.0 kabul edilebilir aralığı içinde yer almasıdır (Linacre, 2014). Başarı testindeki yedi ve sekiz numaralı soruların puanlanmasında kullanılan beşli derecelendirmeye sahip rubrikler bu şartları sağlamaktadır. Testteki ilk altı maddenin puanlanmasında kullanılan dörtlü derecelendirmeye sahip rubriklerde ise, bir kategorisine karşılık gelen ortalama ölçümün sıfır kategorisine

gelen ortalama ölçümden daha düşük olduğu bulunmuştur. Bu bulgu, puanlayıcıların rubrikteki bir ve sıfır kategorilerini birbirinden yeterince iyi ayırt edemediği anlamına gelebilir. Bununla birlikte, uygunluk dışı istatistiklerinin .05 ile 2.0 aralığının (Myford ve Wolfe, 2003) içinde kaldığı dikkate alındığında, puanlayıcıların sıfır ile bir kategorilerinin ayırt edebilme düzeylerinin kabul edilebilir sınırlar içerisinde yer aldığı söylenebilir. Diğer bir ifadeyle, tablodaki uygunluk dışı istatistiklerinin, .5 ile 2.0 ölçütünün içinde yer alması, ortalama ölçüm ile beklenen ölçüm arasındaki farkın kabul edilebilir sınırlardan önemli bir sapma göstermediğini düşündürmektedir. Buna göre, beşli derecelendirmeye sahip rubrikler kadar etkili çalışmasa da; dörtlü derecelendirmeye sahip rubriklerin de kabul edilebilir etkinlikte çalıştığı ifade edilebilir. Nitekim puanlama ölçeği kategorilerinin olasılık eğrilerinde de bu durum görülmektedir. Şekil 4.5'te, dörtlü derecelendirmeye sahip rubrikler için puanlama ölçeği kategorilerinin olasılık eğrisi sunulmuştur. Şekil 4.5 incelendiğinde, puanlama ölçeğinin faklı kategorilerine ait eğrilerin üst üste düşmediği görülmektedir. Her bir eğrinin en yüksek değerine ulaştığı yetenek düzeyi bir diğerinden farklıdır. Ayrıca, rubriğin alt ucundan üst ucunda doğru gidildikçe, kategori eğrilerinin en yüksek değerine ulaştığı yetenek düzeyi artmaktadır. Dolayısıyla, Şekil 4.5'ten hareketle, rubriklerdeki kategorilerin puanlayıcılar tarafından ayırt edilebildiği ve yetenek düzeyi arttıkça rubrikteki üst kategorilerin kullanım olasılığının arttığı söylenebilir.



Şekil 4.5. Başarı testindeki ilk altı madde için SOLO taksonomisine dayalı rubrik kategorilerinin olasılık eğrisi

Testteki yedi ve sekiz numaralı soruların puanlanmasında kullanılan beşli derecelendirmeye sahip rubrikler için puanlama ölçeği kategorilerinin olasılık eğrisi ise Şekil 4.6'da sunulmuştur. Şekil 4.6'ya bakıldığında, rubrikteki kategorilerin her birinin en yüksek değerine ulaştığı yetenek düzeyinin farklı olduğu ve yetenek düzeyi arttıkça puanlama ölçeğinin üst kategorilerinin kullanım olasılığının arttığı görülmektedir. Dolayısıyla, beşli derecelendirmeye sahip SOLO taksonomisine dayalı rubriklerin de etkin bir biçimde çalıştığı söylenebilir.



Şekil 4.6. Başarı testindeki yedi ve sekiz numaralı maddeler için SOLO taksonomisine dayalı rubrik kategorilerinin olasılık eğrisi

4.1.4. Dördüncü Alt Probleme İlişkin Bulgular

Araştırmanın dördüncü alt problemi "SOLO taksonomisine dayalı rubrikler kullanılarak yapılan puanlamalarda, *i)* puanlayıcı katılığı ve cömertliği, *ii)* merkeze yönelme etkisi, *iii)* halo etkisi, *iv)* tutarsızlık ve *v)* yanlılık bulunmakta mıdır?" şeklinde ifade edilmiştir. Çok yüzeyli Rasch analizinden elde edilen çıktılar her bir puanlayıcı etkisine ilişkin literatürde yer alan ölçütler doğrultusunda incelenerek bu alt probleme yanıt aranmıştır. Sıralanan puanlayıcı etkilerine yönelik olarak elde edilen bulgular başlıklar halinde aşağıda sunulmuştur.

4.1.4.1. Puanlayıcı Katılığı ve Cömertliği

Puanlayıcıların katılık ve cömertlikleri yönüyle farklılık gösterip göstermediğini saptamak için ilk olarak grup düzeyindeki istatistiksel göstergeler incelenmiştir. Bu kapsamda, puanlayıcı yüzeyine ilişkin ayırma oranı, ayırma indeksi, güvenirlik indeksi ve Ki Kare değerine bakılmıştır. Puanlayıcı yüzeyine ilişkin ayırma oranı .69, ayırma indeksi 1.25 [(4×.69+1)/3] ve güvenirlik indeksi .32 olarak elde edilmiştir. Bu değerlerin küçük olması, katılık ve cömertlikleri yönüyle puanlayıcılar arasında fark bulunmadığına işaret etmektedir. Ancak bu konuda kesin bir yargıya varmadan önce Ki Kare değerinin incelenmesi gerekmektedir. Ki Kare değerinin anlamlı olmaması [χ^2=8.8, *sd*=6, *p*>.05], puanlayıcıların benzer katılık/cömertlikte puanlama yaptıklarını göstermektedir.

Grup düzeyindeki incelemelerin ardından bireysel düzeydeki istatistiksel göstergeler incelenmiştir. Puanlayıcıların değişken haritası üzerindeki konumları, puanlayıcı katılığı ve cömertliğinin bireysel düzeydeki istatistiksel göstergelerinden biridir. Şekil 4.4'teki değişken haritasına bakıldığında, puanlayıcıların logit ölçeğinin sıfır noktasında kümelendiği görülmektedir. Puanlayıcıların logit ölçeğinin tek bir noktasında kümelenmesi, benzer katılık ve cömertlikte puanlama yaptıklarına işaret etmektedir. Puanlayıcı katılığı ve cömertliğinin bireysel düzeydeki istatistiksel göstergelerinden bir diğeri, araştırmaya dâhil edilen puanlayıcılar için hesaplanan *t* değerleridir. Puanlayıcılara ait logit ölçüleri, bu logit ölçülerinin ortalaması ve standart hatası kullanılarak elde edilen *t* değerleri Tablo 4.16'da sunulmuştur.

Tablo 4.16. SOLO taksonomisine dayalı rubrikler kullanılarak yapılan puanlamalarda katılık ve cömertlikleri açısından puanlayıcılar arasında gözlenen farkın anlamlılığına ilişkin *t*-testi sonuçları

Puanlayıcı	*t* değeri	Farkın Anlamlılığı
P5	1.50	$\lvert t_{hesaplanan}\rvert < t_{kritik}$ olduğundan fark anlamlı değildir.
P4	1.00	
P1	.5	
P6	.5	
P3	-1.00	
P7	-1.25	
P2	-1.50	

Tablo 4.16'ya göre, hesaplanan *t* değerleri -1.50 ile 1.50 arasında değişmektedir. Araştırmaya yedi puanlayıcı dâhil edildiğinden, serbestlik derecesi 7-1=6 olarak hesaplanmış ve altı serbestlik derecesi ile .01 düzeyindeki kritik *t* değeri 3.71 olarak belirlenmiştir. Hesaplanan *t* değerlerinin kritik *t* değerini aşmaması,

katılık ve cömertlikleri açısından puanlayıcılar arasında fark bulunmadığını ortaya koymaktadır.

4.1.4.2. Merkeze Yönelme Etkisi

SOLO taksonomisine dayalı rubriklere göre yapılan puanlamalarda merkeze yönelme etkisinin bulunup bulunmadığını tespit etmek için ilk olarak grup düzeyindeki istatistiksel göstergeler incelenmiştir. Ardından bireysel düzeydeki incelemelere geçilmiştir. Grup düzeyindeki incelemeler, kategori istatistikleri tablosunun gözden geçirilmesiyle başlamıştır. Kategori istatistikleri tablosuna (Tablo 4.15) göre, hem ilk altı maddenin puanlanmasında kullanılan dörtlü derecelendirmeye sahip rubriklerde, hem de yedi ve sekiz numaralı maddelerin puanlanmasında kullanılan beşli derecelendirmeye sahip rubriklerde, tüm kategoriler dengeli bir biçimde kullanılmıştır. Buna göre, SOLO taksonomisi kullanılarak yapılan puanlamalarda merkeze yönelme etkisinin bulunmadığı söylenebilir. Ancak, puanlamalarda merkeze yönelme etkisinin bulunup bulunmadığı konusunda karar vermek için kategori istatistikleri tablosu tek başına yeterli olamamaktadır. Bu durum, merkeze yönelme etkisini belirlemeye yönelik diğer istatistiksel göstergelerin de incelenmesini gerektirmektedir.

Birey yüzeyine ilişkin ayırma oranı, ayırma indeksi, güvenirlik indeksi ve Ki Kare değeri puanlamalarda merkeze yönelme etkisinin bulunup bulunmadığını belirlemek için incelenebilecek grup düzeyindeki diğer istatistiksel göstergelerdir. SOLO taksonomisine göre yapılan puanlamalarda birey yüzeyi için elde edilen ayırma oranı 4.28, ayırma indeksi 6.04 [(4×4.28+1)/3] ve güvenirlik indeksi .95 olarak bulunmuştur. Bu değerlerin yüksek olması, öğrencilerin yetenek düzeyleri açısından ayırt edilebildiğini göstermekte ve puanlamalarda merkeze yönelme etkisinin bulunmadığına işaret etmektedir. Yine, birey yüzeyine ilişkin Ki Kare değerinin anlamlı olması [χ^2=1610.4, *sd*=103 p<.01], yetenek düzeyleri açısından öğrenciler arasında gözlenen farkların manidar olduğunu yansıtmakta ve SOLO taksonomisine göre yapılan puanlamalarda merkeze yönelme etkisinin bulunmadığına işaret eden diğer istatistiksel göstergeleri doğrulamaktadır.

Madde yüzeyine ait uygunluk istatistikleri puanlamalarda merkeze yönelme etkisinin bulunup bulunmadığını tespit etmek için yararlanılabilecek istatistiksel göstergeler arasında yer almaktadır. Uygunluk içi ve uygunluk dışı istatistiklerinin, bu istatistiklerin beklenen değeri olan 1'den önemli ölçüde düşük olması,

puanlamalarda merkeze yönelme etkisinin bulunduğuna yönelik sinyaller şeklinde yorumlanmaktadır. Madde yüzeyine ilişkin ölçüm raporlarına bakıldığında, testteki maddelerden hiç birinin uygunluk içi ve uyguluk dışı istatistiklerinin bu istatistiklerin kabul edilebilir alt sınırı olan .05'in altında yer almadığı saptanmıştır. Bu bulguya göre, SOLO taksonomisine dayalı rubrikler kullanılarak yapılan puanlamalarda merkeze yönelme etkisinin bulunmadığı söylenebilir.

Grup düzeyindeki istatistiksel göstergelerin incelenmesinin ardından bireysel düzeydeki göstergelere bakılmıştır. Her bir puanlayıcıya ilişkin uygunluk istatistiklerine bakılması, bireysel düzeydeki incelemelerin ilk adımını oluşturmuştur. Uygunluk istatistiklerinin kabul edilebilir alt sınır olan .5'in altında olması, puanlayıcıların aşırı tutarlı puanlamalar yaptığını göstermektedir. Aşırı tutarlı puanlamalar ise, puanlamalara merkeze yönelme etkisinin karıştığına işaret etmektedir. SOLO taksonomisine dayalı rubriklere göre yapılan puanlamalarda; puanlayıcılar için hesaplanan uygunluk içi istatistiklerinin .91 ile 1.04 arasında, uygunluk dışı istatistiklerinin ise 1.00 ile 1.19 arasında değiştiği belirlenmiştir. Bu değerler, SOLO taksonomisine dayalı rubriklere göre yapılan puanlamalarda, merkeze yönelme etkisinin bulunmadığı anlamına gelmektedir. Ancak; puanlamalarda merkeze yönelme etkisinin bulunup bulunmadığına dair son karar verilmeden önce, her bir puanlayıcıya ait kategori istatistikleri tablosu incelenmelidir. Başarı testindeki ilk altı maddede, araştırmaya dâhil edilen yedi puanlayıcının her biri için elde edilen kategori istatistikleri, Tablo 4.17'de sunulmuştur.

Tablo 4.17. SOLO taksonomisine dayalı rubrikler kullanılarak puanlanan ilk altı maddede her bir puanlayıcı için hesaplanan kategori istatistikleri

Puanlayıcı	Kategori İstatistikleri			Kalite Kontrol		Uyguluk Dışı İstatistiği
	Kategoriler	Frekans	Yüzde	Ortalama Ölçüm	Beklenen Ölçüm	
P1	0	135	22	-.43	-.62	1.9
	1	111	18	-.38	-.11	.5
	2	153	25	.36	.42	.7
	3	225	36	1.06	1.00	1.2
P2	0	117	19	-.40	-.63	1.6
	1	90	14	-.39	-.10	.7
	2	211	34	.27	.44	1.1
	3	206	33	1.20	1.05	.8
P3	0	138	22	-.40	-.59	1.7
	1	89	14	-.32	-.09	.6
	2	168	27	.34	.44	1.1
	3	229	37	1.11	1.02	1.0
P4	0	139	22	-.43	-.65	1.8
	1	96	15	-.41	-.14	.8
	2	176	28	.27	.39	1.1
	3	213	34	1.03	.97	1.0
P5	0	142	23	-.35	-.68	2.3
	1	109	17	-.51*	-.16	.6
	2	188	30	.30	.39	.9
	3	185	30	1.10	.97	.8
P6	0	137	22	-.38	-.66	2.1
	1	92	15	-.52*	-.14	.5
	2	188	30	.26	.39	.9
	3	207	33	1.08	.98	.9
P7	0	110	18	-.34	-.61	1.7
	1	108	17	-.43*	-.08	.5
	2	184	29	.33	.46	.9
	3	222	36	1.13	1.06	1.0

Tablo 4.17'deki bulgular, puanlama ölçeğinin orta kategorilerini yoğun bir biçimde kullanılan puanlayıcı olmadığını göstermektedir. Buna göre, dörtlü derecelendirmeye sahip SOLO taksonomisine dayalı rubrikler kullanılarak puanlanan ilk altı maddede merkeze yönelme etkisinin bulunmadığı söylenebilir. Ayrıca, Tablo 4.17'ye göre; beş, altı ve yedi numaralı puanlayıcılarda rubriğin bir kategorisine karşılık gelen ortalama ölçümün sıfır kategorisine karşılık gelen ortalama ölçümden daha düşük olduğu belirlenmiştir. Buna göre, söz konusu üç puanlayıcının dörtlü derecelendirmeye sahip SOLO taksonomisine dayalı rubriklerin sıfır ve bir kategorilerini birbirinden yeterince iyi ayırt edemediği ifade edilebilir.

Tablo 4.18. SOLO taksonomisine dayalı rubrikler kullanılarak puanlanan yedi ve sekiz numaralı sorularda her bir puanlayıcı için hesaplanan kategori istatistikleri

Puanlayıcı	Kategori İstatistikleri			Kalite Kontrol		Uyguluk Dışı İstatistiği
	Kategoriler	Frekans	Yüzde	Ortalama Ölçüm	Beklenen Ölçüm	
P1	0	21	10	-.72	-.62	.8
	1	19	9	-.22	-.16	.8
	2	63	30	.06	.17	.8
	3	70	34	.62	.46	.6
	4	35	17	.73	.75	1.0
P2	0	19	9	-.66	-.64	.9
	1	15	7	-.27	-.16	.8
	2	60	29	.07	.19	.7
	3	82	39	.63	.49	.8
	4	32	15	.77	.79	1.0
P3	0	19	9	-.65	-.55	.7
	1	18	9	-.11	-.10	1.0
	2	67	32	.09	.23	.6
	3	54	26	.61	.50	.9
	4	50	24	.81	.79	1.0
P4	0	24	12	-.63	-.64	.9
	1	19	9	-.13	-.18	1.0
	2	60	29	.06	.14	.8
	3	71	34	.52	.42	1.1
	4	34	16	.72	.71	.9
P5	0	20	10	-.74	-.64	.6
	1	15	7	-.42	-.20	.5
	2	57	27	-.04	.12	.8
	3	74	36	.54	.40	.7
	4	42	20	.58	.69	1.3
P6	0	23	11	-.71	-.66	.8
	1	12	6	-.19	-.20	.8
	2	62	30	-.03	.13	.7
	3	78	38	.55	.42	.7
	4	33	16	.75	.71	.9
P7	0	23	11	-.56	-.62	1.2
	1	19	9	-.38	-.15	.8
	2	57	27	.18	.20	.8
	3	79	38	.67	.49	.6
	4	30	4	.77	.79	.8

Başarı testindeki yedi ve sekiz numaralı sorularda, araştırmaya dâhil edilen yedi puanlayıcının her biri için elde edilen kategori istatistikleri, Tablo 4.18'de sunulmuştur. Tablo 4.18'e bakıldığında, tüm puanlayıcıların puanlama ölçeğinin kategorilerini dengeli bir biçimde kullandığı görülmektedir. Buna göre; beşli derecelendirmeye sahip SOLO taksonomisine dayalı rubrikler kullanılarak puanlanan yedi ve sekiz numaralı sorularda da merkeze yönelme etkisinin bulunmadığı söylenebilir. Tablo 4.18'e göre, puanlama ölçeğinin kategorileri arttıkça ortalama ölçümlerin de monoton olarak arttığı belirlenmiştir. Bu bulgu, tüm puanlayıcıların beşli derecelendirmeye sahip SOLO taksonomisine dayalı rubriklerin kategorilerini

birbirinden iyi bir biçimde ayırt edebildiğini yansıtmaktadır. Özetle, gerek grup düzeyindeki gerekse de bireysel düzeydeki istatistiksel göstergeler SOLO taksonomisine dayalı rubrikler kullanılarak yapılan puanlamalarda merkeze yönelme etkisinin bulunmadığını ortaya koymuştur.

4.1.4.3. Halo Etkisi

Halo etkisini belirlemeye yönelik işlemler, grup düzeyindeki ve bireysel düzeydeki istatistiksel göstergelerin incelenmesi olmak üzere iki aşamada gerçekleşmiştir. Madde yüzeyine ilişkin ayırma oranı, ayırma indeksi, güvenirlik indeksi ve Ki Kare değeri halo etkisinin grup düzeyindeki istatistiksel göstergeleridir. SOLO taksonomisine dayalı rubrikler kullanılarak yapılan puanlamalarda madde yüzeyi için hesaplanan ayırma oranı 14.90, ayırma indeksi 20.2 [(4×14.90+1)/3] ve güvenirlik indeksi 1.00 olarak bulunmuştur. Bu değerlerin yüksek olması, maddelerin güçlük düzeyleri açısından farklılık gösterdiğine işaret etmektedir. Hesaplanan Ki Kare değeri de [χ^2=1316.4, *sd*=7, *p*<.05], güçlük düzeyleri açısından maddeler arasında anlamlı fark bulunduğunu yansıtmaktadır. Maddelerin güçlük düzeyleri arasında anlamlı fark olması, öğrencilerin testin farklı maddelerindeki performanslarının birbirinden bağımsız olarak puanlanabildiğini ve puanlamalarda halo etkisinin bulunmadığını göstermektedir.

Linacre'ye (2014) göre, madde yüzeyinin bileşenleri güçlük düzeyleri açısından birbirine eşitlendikten sonra, çok yüzeyli Rasch analizinin tekrarlanması ve puanlayıcılara ait uygunluk istatistiklerinin incelenmesi, puanlamalarda halo etkisinin bulunup bulunmadığı konusunda karar vermek için yapılabilecek bir diğer işlemdir. Maddelerin güçlük düzeyleri açısından eşitlendiği modelde, veri ile mükemmel uyum gösteren puanlayıcıların öğrencilerin testin farklı maddelerindeki performanslarını birbirinden ayırt etmede başarısız olduğu kabul edilmektedir (Myford ve Wolfe, 2004). Bu kapsamda, matematik başarı testindeki sekiz soru güçlük düzeyi açısından birbirine eşitlenerek çok yüzeyli Rasch analizi tekrarlanmıştır. Tekrarlanan analizde; puanlayıcılara ait uygunluk içi istatistiklerinin .94 ile 1.06 arasında değiştiği, uygunluk dışı istatistiklerinin .96 ile 1.08 arasında sıralandığı ve veri ile mükemmel uyum gösteren puanlayıcı bulunmadığı belirlenmiştir. Bu bulgu, SOLO taksonomisine dayalı rubrikler kullanılarak yapılan puanlamalara halo etkisinin karışmadığını göstermektedir.

Puanlamalarda halo etkisinin bulunup bulunmadığının bireysel düzeydeki bir diğer göstergesi, puanlayıcılara ait uygunluk içi ve uygunluk dışı istatistiklerdir. SOLO taksonomisine dayalı rubrikler kullanılarak yapılan puanlamalarda, maddelerin güçlük düzeyleri arasında 1.92 logitlik bir değişim gözlenmiştir. Bu değişime göre, maddelerin güçlük düzeyleri arasındaki farklılaşmanın fazla olduğu söylenebilir. Maddelerin güçlük düzeyleri arasındaki farklılaşma büyük olduğunda, puanlayıcılara ait uygunluk istatistiklerinin 1'den önemli ölçüde yüksek olması halo etkisine işaret etmektedir (Myford ve Wolfe, 2004). SOLO taksonomisine dayalı rubrikler kullanılarak yapılan puanlamalarda, puanlayıcı yüzeyi için hesaplanan uygunluk içi istatistikleri .91 ile 1.04 arasında değişirken, uygunluk dışı istatistikleri 1.00 ile 1.19 arasında sıralanmaktadır. Uygunluk istatistiklerinin 1'e yakın olması, puanlamalarda halo etkisinin bulunmadığı anlamına gelmektedir.

4.1.4.4. Tutarsızlık Etkisi

SOLO taksonomisine dayalı rubrikler kullanılarak yapılan puanlamalarda, tutarsızlık etkisinin bulunup bulunmadığını belirlemeye yönelik işlemlere grup düzeyindeki göstergelerin incelenmesi ile başlanmıştır. Bu kapsamda ilk olarak, birey yüzeyinin bileşenleri yetenek düzeyleri açısından birbirine eşitlenerek çok yüzeyli Rasch analizi yinelenmiş ve puanlayıcılara ait uygunluk istatistikleri incelenmiştir. Yinelenen analizde, puanlayıcılar için hesaplanan uygunluk içi istatistiklerinin .91 ile 1.04 arasında sıralandığı, uygunluk dışı istatistiklerinin .95 ile 1.07 arasında değiştiği ve veri ile mükemmel uyum gösteren puanlayıcı bulunmadığı saptanmıştır. Birey yüzeyinin bileşenlerinin yetenek düzeyi açısından eşitlendiği modelde, veri ile mükemmel uyum gösteren puanlayıcıların tutarsız puanlamalar yaptığı kabul edilmektedir (Linacre 2003'ten akt. Myford ve Wolfe, 2003). Dolayısıyla, araştırmaya dâhil edilen puanlayıcılar arasında SOLO taksonomisine dayalı rubrikleri tutarsız bir biçimde kullanan puanlayıcı olmadığı söylenebilir.

SOLO taksonomisine dayalı rubrikler kullanılarak yapılan puanlamalarda tutarsızlık etkisinin bulunup bulunmadığını belirlemek için incelenen grup düzeyindeki diğer gösteregeler; birey yüzeyine ilişkin ayırma oranı, ayırma indeksi, güvenirlik indeksi ve Ki Kare değeridir. Birey yüzeyi için elde edilen ölçüm raporlarına bakıldığında; ayırma oranı 4.28, ayırma indeksi 6.04 [(4×4.28+1)/3] ve güvenirlik indeksi .95 olarak bulunmuştur. Bu değerlerin yüksek olması öğrencilerin yetenek düzeyleri açısından ayırt edilebildiğini göstermektedir. Aynı şekilde Ki Kare

değerinin anlamlı çıkması [χ^2=1610.4, *sd*=103, *p*<.01], yetenek düzeyleri açısından öğrenciler arasında anlamlı fark bulunduğunu ortaya koymaktadır. Dolayısıyla grup düzeyindeki istatistiksel göstergeler, puanlama ölçeğinin kategorilerini tutarsız bir biçimde kullanan puanlayıcı bulunmadığına yönelik birer kanıt olarak değerlendirilebilir.

SOLO taksonomisine dayalı rubrikler kullanılarak yapılan puanlamalarda tutarsızlık etkisinin bulunup bulunmadığını tespit etmeye yönelik çalışmaların ikinci aşamasında bireysel düzeydeki incelemelere geçilmiştir. Bu doğrultuda öncelikle, puanlayıcılar için rapor edilen uygunluk istatistiklerine bakılmıştır. Herhangi bir puanlayıcı için hesaplanan uygunluk istatistiklerinin 1'den önemli ölçüde büyük olması, söz konusu puanlayıcının puanlama ölçeğinin kategorilerini kullanırken tutarsız davrandığına işaret etmektedir. SOLO taksonomisine dayalı rubrikler kullanılarak yapılan puanlamalarda, puanlayıcılar için hesaplanan uygunluk içi istatistiklerinin .91 ile 1.04 arasında sıralandığı, uygunluk dışı istatistiklerinin ise 1.00 ile 1.19 arasında değiştiği belirlenmiştir. Uygunluk istatistikleri için rapor edilen bu değerler .05 ile 2.0 kabul edilebilir aralığı içinde yer almaktadır (Myford ve Wolfe, 2004). Bu bulgu, puanlama ölçeğinin kategorilerini kullanırken, tutarsız davranan puanlayıcı bulunmadığına işaret etmektedir. Tutarsızlık etkisini belirlemeye yönelik bireysel düzeydeki incelemeler kapsamında yapılabilecek bir diğer işlem, farklı puanlayıcılar tarafından yapılan puanlamalar arasındaki korelasyon (çift serili korelasyon) katsayılarının karşılaştırılmasıdır. Örneğin, *n* puanlayıcının görev aldığı bir araştırmada, birinci puanlayıcının yaptığı puanlamalar ile geriye kalan *n-1* puanlayıcı tarafından yapılan puanlamalar arasındaki korelasyon K_1, ikinci puanlayıcının yaptığı puanlamalar ile geriye kalan *n-1* puanlayıcı tarafından yapılan puanlamalar arasındaki korelasyon K_2…ve *n*. puanlayıcının yaptığı puanlamalar ile geriye kalan *n-1* puanlayıcı tarafından yapılan puanlamalar arasındaki korelasyon K_n olsun. Puanlamalarda tutarsızlık etkisinin bulunup bulunmadığını belirlemek için bu *n* tane korelasyon katsayısı karşılaştırılır. Böylelikle, diğerlerine göre anlamlı derecede daha düşük olan korelasyon katsayıları belirlenir. Belirlenen bu korelasyon katsayılarının karşılık geldiği puanlayıcıların puanlama ölçeğini diğer puanlayıcılara göre daha farklı yorumladığı ve tutarsız puanlamalar yaptığı kabul edilmektedir (Myford ve Wolfe, 2004). SOLO taksonomisine dayalı rubrikler kullanılarak yapılan puanlamalar için hesaplanan korelasyon katsayılarının .43 ile .45 aralığında bulunduğu belirlenmiştir. Hesaplanan

yedi farklı korelasyon katsayısının hiçbiri diğerlerine göre anlamlı derecede daha düşük olmadığından, tutarsız puanlamalar yapan puanlayıcı olmadığı ifade edilebilir. Hung, Chen ve Chen'e göre (2012), hesaplanan korelasyon katsayılarının orta düzeyde olması da puanlama ölçeğinin kategorilerini diğer puanlayıcılara göre daha farklı yorumlayan puanlayıcı olmadığı anlamına gelmektedir. Gerek grup düzeyinde gerekse de bireysel düzeyde yapılan incelemeler, SOLO taksonomisine dayalı rubriklerin kategorilerini tutarsız bir biçimde kullanan puanlayıcı bulunmadığını ortaya koymaktadır.

4.1.4.5. Yanlılık Etkisi

Puanlayıcıların SOLO taksonomisine dayalı rubrik kullanarak yaptıkları puanlamalarda, öğrencilerin bir kısmına diğerlerine göre daha katı ya da daha cömert davranıp davranmadığını belirlemek için puanlayıcı ile birey etkileşimi (*puanlayıcı*×*birey*) incelenmiştir. Puanlayıcıların testteki maddelerin tümünü aynı katılık ya da cömertlikte puanlayıp puanlamadığını tespit etmek için ise puanlayıcı ve madde yüzeyleri arasındaki etkileşime (*puanlayıcı*×*madde*) bakılmıştır. Puanlayıcıların, öğrencilerin bir kısmının bazı maddelerdeki performansına beklenenden daha düşük ya da yüksek puanlar verip vermediğini saptamak için ise puanlayıcı, birey ve madde yüzeyleri arasındaki etkileşim (*puanlayıcı*×*birey*×*madde*) incelenmiştir. Araştırmada puanlayıcı etkileri ile ilgilenildiğinden, bazı öğrencilerin testteki maddelerin bir kısmında beklenenden daha yüksek ya da daha düşük bir performans gösterip göstermediği kapsam dışı tutulmuştur. Dolayısıyla, birey ve madde yüzeyleri arasındaki etkileşim (*birey*×*madde*) araştırmaya dâhil edilmemiştir.

Yanlılık analizi kapsamında ilk olarak *puanlayıcı*×*birey* etkileşimi incelenmiştir. Araştırmada, 104 öğrencinin açık uçlu matematik sorularına verdikleri yanıtlar yedi puanlayıcı tarafından puanlanmıştır. Dolayısıyla puanlayıcı ve birey yüzeyleri arasında 728 (7×104) olası etkileşim bulunmaktadır. Yanlılık analizi sonucunda, söz konusu 728 etkileşim için hesaplanan *t* testi değerlerinin -1.49 ile 1.89 arasında değiştiği saptanmıştır. Hesaplanan tüm *t* değerlerinin ±2 aralığı içinde kalması, 728 etkileşimden hiçbirinin istatistiksel açıdan anlamlı olmadığını göstermiştir. Buna göre, SOLO taksonomisine dayalı rubrikler kullanılarak yapılan puanlamalarda, "*puanlayıcı*×*birey*" yanlılığının olmadığı söylenebilir. Nitekim "*puanlayıcı*×*birey*" yanlılığına ilişkin Ki Kare değeri de istatistiksel açıdan anlamlı bulunmamıştır [χ^2=132.1, *sd*=728, *p*>.05].

Puanlayıcı×*madde* etkileşimine ilişkin elde edilen bulgular Tablo 4.19 ve Tablo 4.20'de sunulmuştur. Puanlayıcı ve madde yüzeyleri arasında 56 [7 (puanlayıcı) × 8 (madde)] olası etkileşim söz konusudur. Puanlayıcı ve madde yüzeyleri arasındaki etkileşimlerin sayısı, puanlayıcı ve birey yüzeyleri arasındaki etkileşim sayısına göre çok daha azdır. Bundan dolayı, puanlayıcı ve madde yüzeyleri arasındaki olası 56 etkileşimin her biri için hesaplanan *t* değerleri Tablo 4.19'da sunulabilmiştir.

Tablo 4.19. SOLO taksonomisine dayalı rubrikler kullanılarak yapılan puanlamalarda puanlayıcı ve madde yüzeyleri arasındaki etkileşime ilişkin *t* değerleri

		Puanlayıcılar						
		P1	P2	P3	P4	P5	P6	P7
Maddeler	M1	-2.4*	0.69	0.03	0.55	1.1	0	0.05
	M2	0.44	1.59	1.19	-0.86	-3.36*	-1.05	2.19*
	M3	1.25	0.38	-1.61	0.8	-2.06*	-0.05	1.31
	M4	0.66	-0.94	-0.12	-0.72	0.45	0.5	0.2
	M5	0.26	-0.13	0.15	0.24	0.13	-0.65	-0.06
	M6	0.23	-0.84	0.23	0.61	0.18	0.61	-1.07
	M7	-0.9	-0.12	0.49	-0.62	2.28*	-0.03	-1.16
	M8	0.54	-0.35	-0.05	-0.07	0.67	0.28	-1.05

*$|t| \geq 2$

Tablo 4.19'da görüldüğü gibi, puanlayıcı ve madde yüzeyleri arasındaki 56 etkileşimden beşi istatistiksel olarak anlamlıdır. Anlamlı olduğu tespit edilen *puanlayıcı*×*madde* etkileşimleri Tablo 4.20'de yer almaktadır. Tablo 4.20'ye göre, yedi numaralı puanlayıcı iki numaralı maddede beklenenden daha yüksek puanlamalar yapmıştır. Beş numaralı puanlayıcının yedi numaralı maddedeki puanlamaları beklenenden daha yüksek; üç ve beş numaralı maddelerdeki puanlamaları ise beklenenden daha düşük bulunmuştur. Bir numaralı puanlayıcının bir numaralı maddede beklenenden daha düşük puanlamalar yaptığı belirlenmiştir. Tablo 4.20'ye göre ayrıca, *puanlayıcı*×*madde* etkileşimine ilişkin Ki Kare değerinin istatistiksel olarak anlamlı olmadığı saptanmıştır [χ^2=56.5, *sd*=56, $p>.05$]. Buna göre, SOLO taksonomisine dayalı rubrikler kullanılarak yapılan puanlamalarda *puanlayıcı*×*madde* yanlılığının olmadığı söylenebilir.

Tablo 4.20. SOLO taksonomisine dayalı rubrikler kullanılarak yapılan puanlamalarda anlamlı olduğu tespit edilen puanlayıcı×madde etkileşimleri

Puanlayıcı	Madde	Gözlenen Puan	Beklenen Puan	Bias (Logit)	Standart Hata	*t*
P7	M2	263	246.60	.33	.15	2.19
P5	M7	276	253.71	.24	.11	2.28
P5	M3	131	151.00	-.21	.10	-2.06
P1	M1	84	106.45	-.27	.11	-2.40
P5	M2	214	240.17	-.39	.12	-3.36

Ki Kare=56.5, *sd*=56, *p*>.05

Yanlılık analizi kapsamında son olarak *puanlayıcı*×*birey*×*madde* etkileşimleri incelenmiş ve elde edilen bulgular Tablo 4.21'de sunulmuştur. Bu üç yüzey arasında 5824 [7 (puanlayıcı)×104 (birey)×8 (madde)] olası etkileşim bulunmaktadır. Yanlılık analizi sonucunda elde edilen *t* değerleri, 5824 etkileşimden, yalnızca 32'sinin istatistiksel olarak anlamlı olduğunu göstermiştir. Anlamlı olduğu tespit edilen *puanlayıcı*×*birey*×*madde* etkileşimleri Tablo 4.21'de sunulmuştur. Puanlayıcı, birey ve madde yüzeyleri arasındaki 5824 etkileşimin tümüne ilişkin analiz çıktılarının Tablo 4.21'de gösterilmesi mümkün olmadığından, tabloda yalnızca üç yüzey arasındaki anlamlı etkileşimlere yer verilmiştir. Tablo 4.21'deki Ki Kare değerinin anlamlı olmaması [χ^2=2893, *sd*=5824, *p*>.05], SOLO taksonomisine dayalı rubrikler kullanılarak yapılan puanlamalarda, *puanlayıcı*×*birey*×*madde* şeklinde bir yanlılığın söz konusu olmadığını göstermektedir.

Tablo 4.21. SOLO taksonomisine dayalı rubrikler kullanılarak yapılan puanlamalarda anlamlı olduğu tespit edilen puanlayıcı×birey×madde etkileşimleri

Puanlayıcı	Birey	Madde	Gözlenen Puan	Beklenen Puan	Bias (Logit)	Standart Hata	*t*
P1	92	M4	3	.13	3.81	1.66	2.29
P5	92	M4	2	.13	2.81	1.06	2.66
P4	92	M4	2	.12	2.79	1.06	2.63
P6	92	M4	2	.12	2.77	1.06	2.61
P3	92	M4	2	.13	2.71	1.06	2.56
P7	92	M4	2	.14	2.70	1.06	2.55
P2	92	M4	2	.14	2.69	1.06	2.54
P4	15	M8	3	.40	2.58	1.19	2.16
P3	23	M1	2	.21	2.35	1.06	2.22
P4	60	M3	2	.21	2.34	1.06	2.21
P2	23	M1	2	.21	2.33	1.06	2.20
P7	43	M4	2	.22	2.29	1.06	2.16
P2	60	M3	2	.23	2.25	1.06	2.12
P2	26	M1	2	.23	2.17	1.06	2.04
P2	87	M8	2	3.46	-1.87	.91	-2.07
P4	87	M7	2	3.49	-1.95	.91	-2.15
P7	87	M7	2	3.53	-2.04	.91	-2.25
P3	37	M8	1	3.34	-2.42	.94	-2.57
P5	57	M5	1	2.77	-2.49	1.01	-2.47
P4	79	M2	1	2.80	-2.66	1.01	-2.64
P5	53	M5	0	2.71	-3.00	1.39	-2.16
P6	53	M5	0	2.72	-3.00	1.35	-2.22
P5	79	M2	0	2.80	-3.68	1.65	-2.23
P1	79	M2	0	2.80	-3.71	1.64	-2.26
P6	79	M2	0	2.81	-3.72	1.64	-2.26
P5	98	M2	0	2.82	-3.78	1.63	-2.32
P2	79	M2	0	2.82	-3.78	1.63	-2.32
P4	98	M2	0	2.82	-3.80	1.63	-2.33
P1	98	M2	0	2.83	-3.81	1.62	-2.35
P6	98	M2	0	2.83	-3.82	1.62	-2.35
P3	98	M2	0	2.84	-3.86	1.62	-2.39
P7	98	M2	0	2.84	-3.87	1.62	-2.39

Ki Kare= 2531.8, *sd*=5824, *p*>.05

4.1.5. Beşinci Alt Probleme İlişkin Bulgular

Araştırmanın beşinci alt problemi; "Standart ve SOLO taksonomisine dayalı rubrik kullanılarak yapılan puanlamalardan elde edilen yetenek kestirimleri ile öğrencilerinin matematik karne notları ve matematik dersi ortak sınavındaki doğru sayıları arasındaki korelasyonlar nasıldır" şeklinde ifade edilmiştir. Bu alt probleme yanıt aramak için standart ve SOLO taksonomisine dayalı rubrikler kullanılarak yapılan puanlamaların FACET programında analiz edilmesi sonucunda elde edilen yetenek kestirimleri bir SPSS dosyasına kaydedilmiştir. Ardından, bu SPSS dosyasına öğrencilerin 2013-2014 Öğretim Yılı Güz Dönemi'ne ait matematik karne notları ile aynı yarıyılda uygulanan matematik dersi ortak sınavındaki doğru sayıları aktarılmıştır. Daha sonra, değişkenler arasındaki ilişki Pearson Momentler Çarpımı

korelasyonu ile analiz edilmiştir. Korelasyon analizi sonucunda elde edilen bulgular Tablo 4.22'de sunulmuştur.

Tablo 4.22. Standart ve SOLO taksonomisine dayalı rubrik kullanılarak yapılan puanlamalardan elde edilen yetenek kestirimleri ile öğrencilerin matematik karne notları ve matematik dersi ortak sınavındaki doğru sayıları arasındaki korelasyonlar

Değişkenler	**1.**	**2.**	**3.**	**4.**
1. Standart rubrik kullanılarak yapılan puanlamalardan elde edilen yetenek kestirimleri	1	.948	.498	.580
2. SOLO taksonomisine dayalı rubrik kullanılarak yapılan puanlamalardan elde edilen yetenek kestirimleri		1	.544	.639
3. Öğrencilerin matematik karne notları			1	.754
4. Öğrencilerin matematik dersi ortak sınavdaki doğru sayıları				1

Tablo 4.22'deki bulgulara göre, standart rubrik kullanılarak yapılan puanlamalardan elde edilen yetenek kestirimleri ile öğrencilerin matematik karne notları [r=.498, p<.01] ve matematik dersi ortak sınavındaki doğru sayıları [r=.580, p<.01] arasında pozitif yönde anlamlı ilişkiler bulunmaktadır. Benzer şekilde, SOLO taksonomisine dayalı rubrik kullanılarak yapılan puanlamalardan elde edilen yetenek kestirimleri ile öğrencilerin matematik karne notları [r=.544, p<.01] ve matematik dersi ortak sınavındaki doğru sayıları [r=.639, p<.01] arasında pozitif anlamlı ilişkiler mevcuttur. Matematik karne notu ve matematik dersi ortak sınavındaki doğru sayısı değişkenleri öğrencilerin matematik başarı puanı olarak kabul edildiği takdirde, SOLO taksonomisine dayalı rubrik kullanılarak hesaplanan yetenek kestirimlerinin standart rubrik kullanılarak elde edilen yetenek kestirimlerine kıyasla öğrencilerin matematik başarısını yordama gücünün daha yüksek olduğu ortaya çıkmaktadır. Diğer değişkenler arasındaki ilişkiler incelendiğinde ise, standart ve SOLO taksonomisine dayalı rubrik kullanılarak yapılan puanlamalardan elde edilen yetenek kestirimleri arasında pozitif yönde, güçlü ve anlamlı bir ilişki tespit edilmiştir [r=.948, p<.01]. Yine Tablo 4.22'ye göre, öğrencilerin matematik karne notları ve matematik dersi ortak sınavındaki doğru sayıları arasında pozitif yönde, güçlü ve anlamlı bir ilişki bulunduğu saptanmıştır [r=.754, p<.01].

4.1.6. Altıncı Alt Probleme İlişkin Bulgular

Araştırmanın altıncı alt problemi, "Standart ve SOLO taksonomisine dayalı rubrik kullanılarak yapılan puanlamalardan elde edilen yetenek kestirimleri arasında

anlamlı fark var mıdır?" şeklinde ifade edilmiştir. Bu alt probleme yanıt aramak için standart ve SOLO taksonomisine dayalı rubrik kullanılarak yapılan puanlamaların FACET programında analiz edilmesi sonucunda elde edilen yetenek kestirimleri bir SPSS dosyasına kaydedilmiştir. Ardından bu yetenek kestirimleri arasındaki farkın anlamlılığı ilişkili örneklem *t*-testi ile analiz edilmiştir. Analiz sonucunda elde edilen bulgular Tablo 4.23'te gösterilmiştir.

Tablo 4.23. Standart ve SOLO taksonomisine dayalı rubrik kullanılarak yapılan puanlamalardan elde edilen yetenek kestirimleri arasındaki farkı gösteren ilişkili örneklem *t*-testi sonuçları

Puanlamada Kullanılan Rubrik	N	$\overline{X}$	ss	sd	*t*	p	Eta Kare (η^2)
Standart Rubrik	104	-.43	.68	103	33.86	.00	.92
SOLO Taksonomisine Dayalı Rubrik	104	.30	.67				

Tablo 4.23'teki bulgulara göre; öğrencilerin matematik performanslarına ilişkin ortalamalar, standart rubrik kullanılarak yapılan puanlamalar için -.43 logit ve SOLO taksonomisi dayalı rubrikler ile yapılan puanlamalar için .30 logit olarak belirlenmiştir. Buna göre, Standart ve SOLO taksonomisine dayalı rubrik kullanılarak yapılan puanlamalardan elde edilen yetenek kestirimleri arasında .73 logitlik bir fark bulunmaktadır. Bu farkın istatistiksel açıdan anlamlı olup olmadığının belirlenmesi için *t* testinden yararlanılmıştır. İlişkili örneklem *t*-testi sonuçları, standart ve SOLO taksonomisine dayalı rubrik kullanılarak yapılan puanlamalardan elde edilen yetenek kestirimleri arasında anlamlı fark bulunduğu ortaya koymaktadır [$t_{(103)}$=33.86, p<.01]. Bu farkın büyüklüğünü ortaya koymak için Eta Kare değeri hesaplanmış ve .92 olarak belirlenmiştir. Eta Kare'nin yorumlanmasında dikkate alınması önerilen ölçütlere göre, standart ve SOLO taksonomisine dayalı rubrik kullanılarak yapılan yetenek kestirimleri arasındaki farkın büyük olduğu söylenebilir.

4.1.7. Yedinci Alt Probleme İlişkin Bulgular

Araştırmanın yedinci alt problemi "Puanlayıcıların standart ve SOLO taksonomisine dayalı rubrikler hakkındaki görüşleri nasıldır?" şeklinde ifade edilmiştir. Bu alt problemi yanıtlamak için ilk olarak, puanlayıcıların SRDA ile STDRDA'da yer alan kapalı uçlu maddelere verdikleri yanıtlarının aritmetik

ortalaması hesaplanmıştır. Daha sonra puanlayıcıların açık uçlu maddelere verdikleri yanıtlar, anketlerde yer alan maddelerin ortaya koyduğu temalar çerçevesinde sınıflandırılıp, gerekli yerlerde doğrudan alıntılara yer verilerek sunulmuştur.

Tablo 4.24. Puanlayıcıların SRDA ile STDRDA'da yer alan kapalı uçlu maddelere verdikleri yanıtlara ilişkin aritmetik ortalama değerleri

Maddeler	Dereceleme	SOLO Taksonomisine Dayalı Rubriğe İlişkin Ortalama	Standart Rubriğe İlişkin Ortalama
Objektifliği (Puanlama işlemi kim tarafından yapılırsa yapılsın aynı sonuca ulaşılması)	[*Çok yüksek (5)* → *Çok düşük (1)*]	4.33	3.00
Ölçülen özellik açısından farklı seviyelerdeki öğrencileri birbirinden ayırt edilebilme özelliği		3.83	3.00
Öğrenciye güçlü ve zayıf olduğu noktalar hakkında geri bildirimler sunabilme özelliği		3.83	3.00
Hazırlanması	[*Çok kolay (5)* → *Çok zor (1)*]	2.67	2.33
Kullanımı		4.50	3.17

Tablo 4.24'te yer alan bulgulara göre, standart rubriklerin objektifliğine ilişkin ortalamanın ($\overline{X}$ =3.00) SOLO taksonomisine dayalı rubriklerin objektifliğine ilişkin ortalamadan ($\overline{X}$ =4.33) daha düşük olduğu belirlenmiştir. Dolayısıyla Tablo 4.24'teki veriler, puanlayıcıların SOLO taksonomisine dayalı rubrikleri standart rubriklere kıyasla daha objektif bulduğunu göstermektedir. Nitekim puanlayıcıların "Standart rubriklerin objektifliği hakkında ne düşünüyorsunuz?" maddesine verdikleri yanıtlar, standart rubriği yeterince objektif bulmadıklarını ortaya koymuştur. Diğer taraftan, puanlayıcıların "SOLO taksonomisine dayalı rubriklerin objektifliği hakkında ne düşünüyorsunuz?" maddesine verdikleri yanıtlar; SOLO taksonomisine dayalı rubrikler ile öğrenci yanıtlarının puanlayıcıdan bağımsız olarak puanlanabildiğini ve bu rubriklerin standart rubriklere kıyasla daha objektif olduğunu göstermiştir. Standart ve SOLO taksonomisine dayalı rubriklerin objektifliği hakkındaki bazı puanlayıcı söylemleri aşağıdaki gibidir.

…standart rubriklerin yeterince objektif olmadığını düşünüyorum. SOLO taksonomisine dayalı rubrikleri ise objektif buluyorum. SOLO taksonomisine dayalı rubrikler ile yaptıkları puanlamalarda çoğu puanlayıcı aynı öğrenci cevabına aynı puanı vermiştir diye düşünüyorum…[P7]

…standart rubrikler soruyu tam yanıtlayan öğrenciler için objektif olabilse de; soruya kısmen cevap veren öğrenciler için tam da objektif değil…[P1]
…SOLO taksonomisine dayalı rubriklerin objektif olduğu kanısındayım. Çünkü değerlendirme işlemini mümkün olduğunca puanlayıcıdan bağımsız hale getiriyor…[P2]

Tablo 4.24'e göre, ölçülen özellik açısından farklı seviyelerdeki öğrencileri birbirinden ayırt edilebilme özelliği açısından SOLO taksonomisine dayalı rubrikler ($\overline{X}$ =3.83) standart rubriklere göre ($\overline{X}$ =3.00) daha etkili bulunmuştur. Puanlayıcılar, SOLO taksonomisine dayalı rubriklerin farklı yetenek düzeylerindeki öğrencilerin birbirinden ayırt edilmesi konusunda etkili olduğu şeklinde görüş bildirmişlerdir. Öte yandan, standart rubriklerin farklı yetenek düzeyindeki öğrencileri yeterince iyi ayırt edemediğini belirtmişlerdir. Bu konuda, puanlayıcılardan bazıları görüşlerini aşağıda şekilde dile getirmiştir.

...SOLO taksonomisine dayalı rubrikler öğrenciler arasındaki farklılığın görülmesinde etkin rol oynamaktadır…[P1]
…Ölçülen özellik açısından farklı yetenek düzeyinde bulunan öğrencilerin ayırt edebilmesi konusunda standart rubriklerin yetersiz olacağını düşünüyorum…[P2]
…Standart rubrikler, alt ve üst düzeydeki öğrencileri birbirinden ayırt edebiliyor ancak orta düzeydeki öğrencileri yeterince iyi ayırt edemiyor…[P4]
… SOLO taksonomisinin öğrencileri yetenek düzeylerine göre ayırt edebileceğini düşünüyorum…[P5]

Tablo 4.24'te görüldüğü üzere, puanlayıcılar öğrencilere güçlü ve zayıf olduğu noktalar hakkında geri bildirimler sunabilme konusunda SOLO taksonomisine dayalı rubriklerin ($\overline{X}$ =3.83) standart rubriklere ($\overline{X}$ =3.00) kıyasla daha etkili olduğu görüşündedirler. Puanlayıcılar, standart rubriklerin hangi öğrenci cevabına kaç puanın verilmesi gerektiği konusunda yeterince açık olmadığını ve bu durumun öğrencilere anlamlı geri dönütler vermeyi zorlaştırdığını ifade etmişlerdir. SOLO taksonomisinin öğrencilere güçlü ve zayıf olduğu noktalar hakkında geri bildirim vermede daha etkili olduğunu dile getiren bir puanlayıcı düşüncelerini şöyle ifade etmiştir.

… SOLO taksonomisine dayalı rubrikleri standart rubrikler ile karşılaştırdığımda dönüt için SOLO taksonomisine dayalı rubriklerin daha kullanışlı olduğunu düşünüyorum. Çünkü SOLO taksonomisine dayalı rubrikler öğrencilerin hangi konularda eksiğinin olduğuna dair direk nokta atışı şeklinde ipuçları veriyor…[P1]

Tablo 4.24'e göre, puanlayıcılar hem SOLO taksonomisine dayalı rubriklerin ($\overline{X}$ =2.67) hem de standart rubriklerin ($\overline{X}$ =2.33) hazırlanmasının zor olduğunu düşünmekle birlikte; SOLO taksonomisine dayalı rubriklerin standart rubriklere göre daha kolay hazırlanabileceğini bildirmişlerdir. SOLO taksonomisine dayalı rubriklerin standart rubriklere göre daha kolay hazırlanabileceğini ifade eden bir puanlayıcı görüşlerini aşağıdaki gibi dile getirmiştir.

...SOLO taksonomisine dayalı rubriklerin hazırlanması zor ve zaman alıcı olsa da standart rubriklere göre daha kolay bir biçimde hazırlayabileceğimi düşünüyorum... [P1]

Tablo 4.24'e göre, SOLO taksonomisine dayalı rubriklerin kullanım kolaylığına ilişkin ortalamanın ($\overline{X}$ =4.50) standart rubriklerin kullanım kolaylığına ilişkin ortalamadan ($\overline{X}$ =3.17) daha yüksek olduğu saptanmıştır. Buna göre, puanlayıcılar SOLO taksonomisine dayalı rubriklerin kullanımının standart rubriklere göre daha kolay olduğu görüşündedirler. Bu durum puanlayıcıların standart ve SOLO taksonomisine dayalı rubriklerin kullanım kolaylığına ilişkin açık uçlu sorulara verdikleri yanıtlarda da görülmektedir. Standart ve SOLO taksonomisine dayalı rubriklerin kullanım kolaylığına ilişkin puanlayıcı görüşlerinden doğrudan alıntılara aşağıda yer verilmiştir.

…SOLO taksonomisine dayalı rubriklerde cevapları puanlama daha rahat. Standart rubriklere göre karmaşıklığı ortadan kaldırıyor...[P4]

…SOLO taksonomisine dayalı rubrikte en beğendiğim nokta puanlamanın anlaşılır ve kullanımının kolay olmasıydı. Çünkü neye kaç puan verileceği açıktı…[P5]

…SOLO taksonomisine dayalı rubriklerde soruların cevaplarına uygulanacak ölçütler daha anlaşılır bu da kullanımını kolaylaştırıyor... [P7]

Puanlayıcılara standart ve SOLO taksonomisine dayalı rubrikler hakkında yöneltilen bir diğer açık uçlu madde; "Standart/SOLO taksonomisine dayalı rubrikler kullanılmadan önce bu rubriklere yönelik bir eğitim verilmesinin gerekliliği hakkında ne düşünüyorsunuz?" şeklindedir. Puanlayıcılar bu maddeye verdikleri yanıtlarda, hangi rubrik türü kullanılırsa kullanılsın puanlama işleminden önce puanlamada kullanılacak rubriklere yönelik bir eğitim verilmesi gerektiğini belirtmişlerdir.

…Bence faydalı olabilir. Çünkü standart rubrikle kâğıt okumak yaygın değil. Bunları bilen ve uygulayan öğretmen sayısı eminim ki azdır…[P2]

…Ben hangi konu olursa olsun yenilik getirecek tüm konularda eğitim verilmesi gerektiğini düşünmekteyim. Örneğin, ben iki yıldır öğretmenim ama sadece kendi yaptığım puanlama anahtarına göre puanlama yapıyordum ve bu sekiz soruyu puanlarken zorlandım açıkçası. Bu yüzden rubrikler hakkında eğitim verilmesi gerektiği taraftarıyım…[P5]

...Eğitimin gerekli olduğunu düşünüyorum. Aksi halde SOLO taksonomisinin ne olduğunu bilmeden bunu doğruca rubriklerde kullansaydık yani uygulamaya dökseydik yanlış sonuçlar elde edebilirdik…[P7]

Puanlayıcıların SRDA ile STDRDA'da yer alan kapalı ve açık uçlu maddelere verdikleri yanıtlar, SOLO taksonomisine dayalı rubrikleri standart rubriklere kıyasla objektiflik, ayrıt edicilik, etkili geri bildirim sunabilme, hazırlama ve kullanım kolaylığı gibi birçok açıdan daha etkili bulduğunu ortaya koymuştur. Bu durum, puanlayıcıların "Açık uçlu matematik soruların puanlanmasında standart ve SOLO taksonomisine dayalı rubriklerden hangisini tercih edersiniz. Neden?" şeklinde ifade edilen açık uçlu maddeye verdikleri yanıtlarda da açık bir biçimde görülmektedir. Bu maddeye ilişkin bazı puanlayıcı cevapları aşağıda sunulmuştur.

…Kesinlikle SOLO taksonomisine dayalı rubriği tercih ederim. Puanlama daha kolay, daha objektif olduğundan…[P4]

…SOLO taksonomisine dayalı rubrikleri tercih ederim. Çünkü daha kullanışlı ve daha objektif olduğu kanısındayım…[P5]

…SOLO taksonomisine dayalı rubrikleri tercih ederim. Nedenleri: Bilenle bilmeyeni ayırt etme konusunda daha uygun, puanlaması bakımından daha kullanışlı, farklı yetenek düzeylerini belirlemede daha etkili…[P7]

4.2. TARTIŞMA

Araştırmadan elde edilen bulgulara ilişkin tartışma, alt problemlerin sırasına uygun olarak aşağıda sunulmuştur.

4.2.1. Birinci Alt Probleme İlişkin Tartışma

Araştırmanın birinci alt problemine ilişkin bulgular; standart rubrikler kullanılarak puanlanan açık uçlu matematik sorularında puanlayıcı, birey ve madde yüzeyleri için hesaplanan güvenirlik indeksinin yüksek ve Ki Kare değerinin anlamlı olduğunu göstermiştir. Bu bulguya dayanarak, farklı yetenek düzeyindeki öğrencilerin birbirinden ayırt edilebildiği, maddelerin güçlük düzeyleri açısından

farklılık gösterdiği, katılık ve cömertlikleri yönüyle puanlayıcılar arasında anlamlı fark olduğu söylenebilir. Yine bu alt probleme ilişkin bulgular, puanlayıcılar arası mutlak uyumun ve puanlayıcılar arası güvenirliğe ilişkin kappa istatistiğinin düşük olduğunu ortaya koymuştur. Performansa dayalı değerlendirmelerde rubrik kullanımının temel amaçlarından biri, puanlama işleminin ne zaman ve kim tarafından yapıldığından bağımsız olarak gerçekleştirilmesini sağlamaktır (Moskal ve Leydens, 2000; Purpura, 2004). Diğer bir ifadeyle, açık uçlu soruların puanlanmasında kullanılan rubriklerin puanlayıcı etkilerini minimum düzeyde tutarak puanlayıcı içi ve puanlayıcılar arası güvenirliği arttırması beklenmektedir (Dunbar, Brooks ve Miller, 2006). Ancak, araştırmadan elde edilen bulgular standart rubriklerin bu beklentiye yeterince cevap veremediğini yansıtmaktadır. Bu bulgu, Güler (2008) tarafından yapılan araştırmanın sonuçlarıyla paralellik göstermektedir. Güler'in (2008) yaptığı çalışmada, öğrencilerin açık uçlu matematik sorularına verdikleri yanıtlar; herhangi bir taksonomi temele alınmadan hazırlanan holistik rubrikler kullanılarak dört farklı puanlayıcı tarafından puanlanmıştır. Araştırma sonucunda; öğrenciler ile maddelerin yüksek güvenirlikte birbirinden ayırt edilebildiği, puanlayıcıların katılık ve cömertlikleri arasında anlamlı fark olduğu sonucuna ulaşılmıştır. Buna göre, standart rubriklerin puanlayıcılar arası güvenirliği arttırmada yeterince işlevsel olmadığı yönündeki araştırma bulgusunun Güler (2008) tarafından yapılan çalışmanın sonuçlarıyla desteklendiği söylenebilir.

Araştırmada katılık ve cömertlikleri açısından puanlayıcılar arasında anlamlı fark olmasına rağmen, birey ve madde yüzeyleri için hesaplanan güvenirlik katsayılarının yüksek olması, analizlerin çok yüzeyli Rasch modeline göre yapılmasından kaynaklanıyor olabilir. Çünkü çok yüzeyli Rasch analizleri, puanlayıcılar arasındaki farklılıkların tespit edilmesi ile sınırlı kalmamaktadır. Aynı zamanda bu farklılıklara yönelik düzeltme işlemi de uygulamaktadır (Abu Kassim, 2007). Bu noktadan hareketle, objektif olarak puanlama yapmanın mümkün olmadığı değerlendirmelerde, çok yüzeyli Rasch analizinin öncelikli olarak tercih edilmesi gereken bir model olduğu söylenebilir.

4.2.2. İkinci Alt Probleme İlişkin Tartışma

Araştırmanın ikinci alt problemin ilişkin bulgular; çalışmaya dâhil edilen yedi puanlayıcıdan altısı için puanlayıcı katılığı ya da cömertliğinin söz konusu olduğunu göstermiştir. Buna göre, puanlayıcı katılığı ve cömertliğinin araştırmaya katılan

puanlayıcıların neredeyse tamamında gözlenen bir puanlayıcı etkisi olduğu ifade edilebilir. Bu bulgu, puanlayıcı katılığı ve cömertliğinin puanlama sürecine karışan en önemli puanlayıcı etkisi olduğu (Cronbach, 1990) şeklindeki kuramsal bilgiler ile desteklenmektedir. Puanlayıcıların katılık ve cömertlikleri arasında tespit edilen anlamlı fark, standart rubriklerin puanlayıcılar arasındaki farklılıkları gidermede yeterince etkili olmadığını düşündürmektedir. Literatüre bakıldığında, standart rubriklerin puanlayıcıların katılık ve cömertlikleri arasındaki farkı ortadan kaldırma konusunda etkili olmadığı şeklindeki bu düşünceyi doğrulayan araştırmaların (Güler, 2008, Güler ve Gelbal, 2010) bulunduğu görülmektedir.

Araştırmaya dâhil edilen bütün puanlayıcılar ve testteki tüm sorular için geçerli olmasa da; standart rubriklere göre yapılan puanlamalara karışan bir diğer puanlayıcı etkisi merkeze yönelmedir. Testteki yedi ve sekiz numaralı soruların puanlanmasında yedi puanlayıcıdan dördü için merkeze yönelme etkisi saptanmıştır. Bu iki soru, testin alt amaçlardan oluşan maddelerdir. Dolayısıyla standart rubriklerin, birden fazla alt amaçtan oluşan maddelerde merkeze yönelme etkisini önleme açısından yeterince etkili olmadığı söylenebilir. Bu durum puanlayıcı görüşlerine de yansımıştır. Puanlayıcılar, standart rubrikleri kullanırken en fazla alt amaçlardan oluşan soruların puanlanmasında zorlandıklarını ifade etmişlerdir. Yaşadıkları bu zorluk, puanlayıcıları puanlama ölçeğinin orta kategorisine başvurmaya yöneltmiş olabilir. Puanlama ölçeğinin orta kategorisinin yoğun olarak kullanılması ise, puanlamalara merkeze yönelme etkisinin karışmasıyla sonuçlanmıştır.

Standart rubriklere göre yapılan puanlamalarda halo etkisinin bulunmadığı belirlenmiştir. Bu bulgu, puanlayıcıların sınav kâğıtlarını puanlarken, öğrenciler üzerinden ilerlemek yerine; maddeler üzerinden ilerlemesinin bir sonucu olabilir. Diğer bir ifadeyle, puanlayıcılar sınav kâğıtlarını değerlendirirken bir öğrencinin testteki tüm maddelere verdiği yanıtları arka arkaya puanlamamıştır. Bunun yerine; ilk olarak tüm öğrencilerin testin birinci sorusuna verdikleri yanıtların puanlanması, daha sonra tüm öğrencilerin testin ikinci sorusuna verdikleri yanıtların puanlanması şeklinde bir yol izlemişlerdir. Puanlamada bu şekilde bir yol takip edilmesinin halo etkisini önlediği düşünülmektedir. Nitekim Pulakos (1991), puanlamada izlenecek böyle bir yaklaşımın halo etkisini azaltma konusunda en etkili yöntemlerden biri olduğunu ifade etmektedir.

Standart rubriklere göre yapılan puanlamalarda tutarsızlık etkisinin bulunmadığı saptanmıştır. Buna göre, puanlama ölçeğinin kategorilerinin tüm puanlayıcılar tarafından aynı şekilde yorumlandığı söylenebilir. Bu sonuç; puanlama işleminden önce uygulanan puanlayıcı eğitimlerinin ve bu eğitimler kapsamında yaptırılan örnek uygulamaların, puanlayıcıların rubrik kategorilerine ilişkin ortak bir algı geliştirmesini kolaylaştırıcı etkisiyle (Alderson, Clapham ve Wall, 1995) açıklanabilir. Puanlayıcı katılığı ve cömertliği, merkeze yönelme etkisi, ranj sınırlaması ve halo etkisi puanlamalardaki sistematik hataları gösteren tek biçimli (uniform) etkiler iken; tutarsızlık puanlamalarda tesadüfi değişkenliklere yol açan ve tek biçimli olmayan bir etkidir (Ramineni, 2008). Puanlamalarda tutarsızlık etkisinin bulunmaması, standart rubriklere göre yapılan puanlamalara tesadüfi değişkenliklerin karışmadığı anlamına gelmektedir. Dolayısıyla, açık uçlu matematik sorularının puanlanmasında kullanılan standart rubriklerin, puanlamalara tesadüfi hataların karışmasını önleme konusunda etkili olduğu ifade edilebilir.

Standart rubriklere göre yapılan puanlamalarda, bulunup bulunmadığı araştırılan bir diğer puanlayıcı etkisi yanlılıktır. Yanlılık analizine ilişkin bulgular incelendiğinde, *puanlayıcı*×*birey*×*madde* ve *puanlayıcı*×*birey* yanlılıklarının istatistiksel açıdan anlamlı olmadığı belirlenmiştir. Dolayısıyla puanlayıcıların tüm öğrencileri aynı katılık/cömertlikte puanladığı söylenebilir. Yine bu sonuç puanlayıcıların bazı öğrencilerin bazı maddelerdeki performanslarını puanlarken beklenenden daha katı ya da daha cömert davranmadığını yansıtmaktadır. Yanlılık; cinsiyet, yaş, kültürel unsurlar gibi çeşitli özelliklerine bağlı olarak puanlayıcıların bazı öğrencilere diğerlerine göre daha yüksek ya da daha düşük puanlar verme eğiliminde olması şeklinde tanımlanan bir puanlayıcı etkisidir (Kumar, 2005). Bu tanım göz önünde bulundurulduğunda, standart rubriklere göre yapılan puanlamalarda *puanlayıcı*×*birey*×*madde* ya da *puanlayıcı*×*birey* yanlılıklarının olmaması, öğrencilerin daha önceki matematik başarılarının ya da diğer demografik özelliklerinin puanlayıcılar tarafından bilinmemesi ile açıklanabilir. Standart rubriklere göre yapılan puanlamalarda, *puanlayıcı*×*madde* yanlılığının ise istatistiksel açıdan anlamlı olduğu sonucuna ulaşılmıştır. Bu sonuç puanlayıcıların testteki tüm maddeleri aynı katılık/cömertlikte puanlamadığını ortaya koymaktadır. Hung, Chen ve Chen'e (2012) göre; puanlayıcıların bazı maddeleri puanlarken beklenenden daha yüksek ya da daha düşük puanlamalar yapması, puanlama ölçeğinin kategorilerini tam olarak anlamamaları ile açıklanabilir. Diğer bir deyişle

puanlayıcıların standart rubrik kullanarak yaptıkları puanlamalarda, hangi cevaba kaç puan vermeleri gerektiğini tam olarak kestirememeleri *puanlayıcı×madde* yanlılığına kaynaklık eden bir faktördür. Standart rubriklerin, öğrencinin cevabına karşılık gelen derecenin belirlenmesi konusunda yeterince açık olmadığı puanlayıcı görüşlerine de yansımıştır. Puanlayıcılar, puanlama sırasında öğrencinin cevabını en iyi yansıtan rubrik kategorisinin hangisi olduğu konusunda tereddütler yaşadıklarını dile getirmişlerdir.

4.2.3. Üçüncü Alt Probleme İlişkin Tartışma

Araştırmanın üçüncü alt probleminde; SOLO taksonomisine dayalı rubrikler kullanılarak puanlanan açık uçlu matematik sorularında puanlayıcı, birey ve madde yüzeyleri için hesaplanan güvenirlik indeksinin yüksek ve Ki Kare değerinin anlamlı olduğu belirlenmiştir. Puanlayıcı yüzeyi için hesaplanan ayırma oranı ve güvenirlik indeksinin ise düşük olduğu tespit edilmiştir. Ayrıca puanlayıcı yüzeyine ait Ki Kare değerinin istatistiksel açıdan anlamlı olmadığı saptanmıştır. Bu bulgulara göre, SOLO taksonomisine dayalı rubriklerin yetenek düzeyleri farklı olan öğrencilerin birbirinden ayırt edilmesine imkân tanıyan bir puanlama ölçeği olduğu söylenebilir. Yine bu bulgulara göre, SOLO taksonomisine dayalı rubriklere göre yapılan puanlamalarda güçlük düzeyleri farklı olan maddelerin yüksek güvenirlikte birbirinden ayırt edilebildiği, katılık ve cömertlikleri yönüyle puanlayıcılar arasında anlamlı fark olmadığı ifade edilebilir. Bu alt probleme yanıt aramak için gerçekleştirilen istatistiksel analizlerin ortaya koyduğu bir diğer bulgu, puanlayıcılar arası mutlak uyumun ve puanlayıcılar arası güvenirliğe ilişkin kappa istatistiğinin yüksek olduğudur. Bu sonuç; SOLO taksonomisine dayalı rubriklerin puanlayıcılar arasındaki farklılıkların giderilmesine yardımcı olduğunu ve puanlama işleminin daha objektif bir biçimde gerçekleşmesini sağladığını düşündürmektedir. Literatürde, SOLO taksonomisine dayalı rubriklerin puanlayıcılar arası güvenirliği arttırdığı yönündeki bu bulguyu destekleyen araştırmalar olduğu gibi; bu bulgu ile çelişen çalışmalar da bulunmaktadır. Örneğin; Burnett (1998) ve Hundzynski (2008) tarafından yapılan araştırmalarda, SOLO taksonomisi kullanılarak gerçekleştirilen değerlendirmelerde puanlayıcılar arası güvenirlik incelenmiştir. Bu araştırmalardan elde edilen puanlayıcılar arası güvenirlik katsayıları sırasıyla .85 ve .87 olarak bulunmuştur. Dolayısıyla, Burnett (1998) ve Hundzynski (2008) tarafından yapılan

çalışmalardan elde edilen bulguların bu araştırmanın sonuçlarıyla paralellik gösterdiği söylenebilir.

Araştırma sonuçlarıyla paralellik gösteren bir başka çalışma Yazıcı (2013) tarafından yapılmıştır. Yazıcı (2013) tarafından yapılan çalışma kapsamında, açık uçlu fizik soruları üç puanlayıcı tarafından SOLO taksonomisine dayalı rubrikler kullanılarak puanlanmıştır. Puanlamadan elde edilen veriler puanlayıcılar arası güvenirliğin yüksek olduğunu ve SOLO taksonomisine dayalı rubriklerin puanlayıcılar arasındaki farklılıkları azalttığını ortaya koymuştur. Leung (2000) tarafından yapılan çalışmada ise, bu araştırmanın bulgularından farklılık gösteren sonuçlara ulaşılmıştır. Leung (2000) tarafından SOLO taksonomisine dayalı rubrik kullanılarak yapılan puanlamalarda, puanlayıcı içi ve puanlayıcılar arası güvenirlik değerlerini incelenmiştir. Araştırmada SOLO taksonomisine dayalı rubrik kullanılarak yapılan puanlama işlemi için hesaplanan puanlayıcı içi ve puanlayıcılar arası güvenirlik katsayıları sırasıyla .71 ve .49 olarak belirlenmiştir. Bu sonuç, SOLO taksonomisine dayalı rubrik kullanılarak yapılan değerlendirmelerde puanlayıcı içi güvenirliğin kabul edilebilir sınırlar içerisinde yer aldığı; ancak puanlayıcılar arası güvenirliğin düşük olduğu şeklinde yorumlanmıştır. Leung'a (2000) göre, SOLO taksonomisine dayalı rubrik kullanılarak yapılan değerlendirmelerde puanlayıcı güvenirliğinin düşük olması; puanlayıcıların SOLO taksonomisine dayalı rubrik ile puanlama yapmaya aşina olmamasından kaynaklanmaktadır. Dolayısıyla Leung (2000) tarafından yapılan çalışma ile bu araştırmada puanlayıcılar arası güvenirliğe ilişkin çelişkili sonuçlar elde edilmesi, bu araştırmada yapılan puanlamalar öncesinde puanlayıcıların SOLO taksonomisine dayalı rubriklerin kullanımına yönelik bir eğitim almasıyla açıklanabilir.

4.2.4. Dördüncü Alt Probleme İlişkin Tartışma

Araştırmanın dördüncü alt problemine ilişkin bulgular, SOLO taksonomisine dayalı rubriklere göre yapılan puanlamalarda; puanlayıcı katılığı ve cömertliği, merkeze yönelme etkisi, halo etkisi, tutarsızlık ve yanlılığın bulunmadığını göstermiştir. Bu bulguya göre, SOLO taksonomisine dayalı rubriklerin tüm puanlayıcılar tarafından aynı şekilde yorumlandığı ve puanlayıcıların rubrikteki kategorileri tam olarak anladığı düşünülebilir. Puanlayıcıların STDRDA'da yer alan açık uçlu sorulara verdikleri yanıtlar bu düşünceyi doğrulamaktadır. Puanlayıcılar, SOLO taksonomisine dayalı rubriklerin açık ve anlaşılır kategorilerinden oluştuğunu

belirtmişlerdir. Bu alt probleme ilişkin bulgular; SOLO taksonomisinin değerlendirmede kullanılacak ölçütleri oldukça açık bir hale getirdiği (Hattie ve Purdie, 1998) ve kolaylıkla anlaşılabilecek düzeylerden oluştuğu (Biggs ve Collis, 1982) şeklindeki kuramsal bilgiler ile örtüşmektedir. SOLO taksonomisine dayalı rubrikler kullanılarak yapılan puanlamalarda sıralanan puanlayıcı etkilerinin bulunmaması, puanlama ölçeğinin kategorilerinin açık olmasının yanı sıra puanlama sürecinde izlenen yol ile de ilişkili olabilir. Söz gelimi, puanlayıcıların puanladıkları öğrencilerin cinsiyet, yaş ve geçmiş matematik başarısı gibi demografik özelliklerini bilmemesi; puanlama işlemine *puanlayıcı×birey* ve *puanlayıcı×birey×madde* yanlılıklarının karışması olasılığını azaltmış olabilir. Puanlamalarda tutarsızlık etkisinin bulunmaması puanlama işleminden önce uygulanan puanlayıcı eğitimlerine bağlanabilir. Bu eğitimler kapsamında puanlayıcılara yaptırılan örnek uygulamalar, rubrik kategorilerinin puanlayıcılar tarafından aynı şekilde anlaşılmasını kolaylaştırmaktadır (Alderson, Clapham ve Wall, 1995). Uygulanan eğitimlerin puanlayıcıların rubrik kategorilerine ilişkin ortak bir algı geliştirmesini kolaylaştıran bu etkisi puanlamalara tutarsızlık etkisinin karışmasını önlemiş olabilir. Puanlama işlemine halo etkisinin karışmaması ise, puanlayıcıların herhangi bir öğrencinin farklı maddelere verdiği yanıtları arka arkaya puanlanmaması ile açıklanabilir. Puanlayıcılar, standart rubriklere göre yapılan puanlamalarda olduğu gibi SOLO taksonomisine dayalı rubrikleri kullanarak yaptıkları puanlamalarda da öğrenciler üzerinden ilerlemek yerine; maddeler üzerinden ilerlemişlerdir. Yani puanlayıcılar ilk olarak tüm öğrencilerin testin birinci sorusuna verdikleri yanıtların puanlanması, daha sonra tüm öğrencilerin testin ikinci sorusuna verdikleri cevapların puanlanması şeklinde bir yol takip etmişlerdir. Puanlamada bu şekilde bir yol takip edilmesi, SOLO taksonomisine dayalı rubrikler kullanılarak yapılan puanlamalara halo etkisinin karışma ihtimalini azalmaktadır.

4.2.5. Beşinci Alt Probleme İlişkin Tartışma

Araştırmanın beşinci alt probleminden elde edilen bulgular; SOLO taksonomisine dayalı rubrik kullanılarak hesaplanan yetenek kestirimlerinin standart rubrik kullanılarak elde edilen yetenek kestirimlerine kıyasla öğrencilerin matematik başarısını yordama gücünün daha yüksek olduğunu göstermiştir. Cebirsel düşünme, istatistiksel düşünme ve geometrik düşünme gibi matematiksel düşünme kapsamında yer alan farklı düşünme biçimlerinin SOLO taksonomisinin düzeyleri ile paralellik

göstermesi bu sonuca kaynaklık eden faktörlerden biri olabilir. Örneğin, Jurdak (1991) tarafından yapılan çalışmada, SOLO taksonomisi ile Van Hiele geometrik düşünme düzeyleri arasında büyük ölçüde benzerlik olduğu belirlenmiş ve geometri dersi öğrenme çıktılarının değerlendirilmesinde SOLO taksonomisinden yararlanılabileceği sonucuna ulaşılmıştır. Mooney (2002) tarafından yapılan çalışmada, öğrencilerin istatistiksel düşünme süreçlerinin ölçülmesinde SOLO taksonomisinin uygun bir model olduğu tespit edilmiştir. Lian ve Idris (2006) ile Lian, Yew ve Idris (2009) tarafından yapılan araştırmalarda ise, öğrencilerin cebir problemlerini çözme becerilerinin değerlendirilmesinde SOLO taksonomisinden yararlanılabileceği saptanmıştır. SOLO taksonomisinin matematik dersi öğrenme çıktılarının değerlendirilmesinde öncelikli olarak tercih edilebilecek bir model olduğunu gösteren söz konusu araştırmalar, bu alt probleme ilişkin bulguların literatür ile desteklendiğine işaret etmektedir.

Araştırmada ayrıca, standart ve SOLO taksonomisine dayalı rubriklere göre yapılan puanlamaların analiz edilmesiyle elde edilen yetenek kestirimleri arasında pozitif yönde ve güçlü bir ilişki olduğu belirlenmiştir. Bu bulguya dayanarak, her iki rubrik türüne göre hesaplanan yetenek kestirimleri arasındaki göreli uyumun oldukça yüksek olduğu söylenebilir. Bir başka ifadeyle, standart ve SOLO taksonomisine dayalı rubrikler kullanılarak elde edilen yetenek kestirimleri referans alınarak öğrenciler arasında bir sıralama yapıldığında, sonuçlar büyük ölçüde eş değer olmaktadır. Dolayısıyla, standart ve SOLO taksonomisine dayalı rubrik kullanılarak yapılan puanlamalar bağıl değerlendirme amacıyla kullanılacak ise, her iki rubrik türüne göre alınacak kararların büyük ölçüde benzer olacağı söylenebilir.

4.2.6. Altıncı Alt Probleme İlişkin Tartışma

Araştırmanın altıncı alt problemine ilişkin bulgular; standart ve SOLO taksonomisine dayalı rubrik kullanılarak yapılan puanlamalardan elde edilen yetenek kestirimleri arasında anlamlı fark olduğunu göstermiştir. Buna göre, standart ve SOLO taksonomisine dayalı rubrik kullanılarak yapılan puanlamalar mutlak değerlendirme amacıyla kullanılacak ise, her iki rubrik türüne göre alınacak kararların farklılık göstereceği söylenebilir.

Bu alt probleme ilişkin bulgular, beşinci alt probleme ilişkin bulgular ile bir arada ele alındığı takdirde, ortaya çıkan sonuçları anlamlandırmanın daha kolay olacağı düşünülmektedir. Araştırmanın beşinci alt probleminde, standart ve SOLO

taksonomisine dayalı rubriklere göre hesaplanan yetenek kestirimleri arasındaki göreli uyumun oldukça yüksek olduğu belirlenmiştir. Bu alt probleme ilişkin bulgular ise, iki rubrik türüne göre elde edilen yetenek kestirimleri arasındaki mutlak uyumun düşük olduğuna işaret etmektedir. Dolayısıyla, açık uçlu matematik sorularının puanlanmasında SOLO taksonomisinin temele alınmasının bağıl değerlendirmeden çok mutlak değerlendirme sonuçlarını etkileyeceği ifade edilebilir. Farklı rubrik türleri kullanılarak yapılan puanlamaların öğrenciler arasındaki göreli sıralamadan çok; yetenek kestirimlerine ilişkin mutlak kararları etkilediğini ortaya koyan bu sonuç, Chi (2001) tarafından yapılan araştırmanın bulgularıyla benzerlik göstermektedir. Chi (2001) tarafından yapılan çalışmada, analitik ve holistik rubrik kullanılarak yapılan puanlamaların performans değerlendirmede oluşturduğu farklılıklar incelenmiştir. Araştırmada sonucunda, öğrenciler arasındaki göreli sıralamanın kullanılan rubrik türünden etkilenmediği belirlenmiştir. Ancak, öğrencilerin yetenek düzeylerine ilişkin mutlak kararların kullanılan rubrik türüne göre farklılık gösterdiği saptanmıştır.

4.2.7. Yedinci Alt Probleme İlişkin Tartışma

Araştırmanın yedinci alt probleminde, puanlayıcıların standart ve SOLO taksonomisine dayalı rubrikler hakkındaki görüşlerinin nasıl olduğu incelenmiştir. Puanlayıcılar, SOLO taksonomisine dayalı rubrikleri standart rubriklere göre daha objektif bulduklarını ifade etmişlerdir. SOLO taksonomisine dayalı rubriklerin değerlendirme sürecinde kullanılabilecek objektif ölçütler sağladığı (Maddrell, 2011) şeklindeki kuramsal bilgiler, puanlayıcıların SOLO taksonomisine dayalı rubriklerin objektifliği hakkındaki görüşleri ile örtüşmektedir. SOLO taksonomisine dayalı rubriklerin objektifliğine ilişkin ampirik araştırmalar (Yazıcı, 2013) da bu bulguyu desteklemektedir. Yazıcı (2013) tarafından yapılan araştırmada, lise öğrencilerinin açık uçlu fizik sorularına verdikleri yanıtlar üç farklı puanlayıcı tarafından puanlanmıştır. Puanlayıcılar, ilk olarak kendi hazırladıkları puanlama anahtarlarını kullanarak puanlama yapmıştır. Daha sonra, aynı öğrenci cevaplarını SOLO taksonomisine dayalı rubriklerden yararlanarak puanlamışlardır. Puanlamaların ardından puanlayıcıların SOLO taksonomisine dayalı rubrikler hakkındaki görüşleri alınmıştır. Her üç puanlayıcı da SOLO taksonomisine dayalı rubriklerin kendi hazırladıkları puanlama anahtarlarından daha objektif olduğunu belirtmiştir.

Puanlayıcılar SOLO taksonomisine dayalı rubrikler hakkındaki görüşlerini açıklarken, bu rubriklerin hazırlanmasının ve kullanımının standart rubriklere göre daha kolay olduğunu dile getirmişlerdir. Yazıcı (2013) tarafından yapılan araştırmadan elde edilen sonuçlar da bu doğrultudadır. Yazıcı (2013) tarafından yapılan araştırmada, puanlayıcılar hazırlanması ve kullanışlılığı açısından SOLO taksonomisine dayalı rubrikleri kendi hazırladıkları puanlama anahtarlarına kıyasla daha etkili bulmuşlardır. Dolayısıyla, SOLO taksonomisine dayalı rubriklerin objektifliğine, hazırlanma ve kullanım kolaylığına ilişkin bu araştırmadan elde edilen bulgular ile Yazıcı (2013) tarafından yapılan çalışmadan elde edilen bulguların benzerlik gösterdiği söylenebilir. Ancak bu benzerlik yorumlanırken, SOLO taksonomisine dayalı rubriklerin sıralanan özelliklerinin bu araştırmada standart rubrikler ile Yazıcı (2013) tarafından yapılan çalışmada ise puanlayıcıların kendi hazırladıkları puanlama anahtarlarıyla karşılaştırılarak değerlendirildiği göz ardı edilmemelidir.

Son olarak, puanlayıcılar ölçülen özellik açısından farklı seviyelerdeki öğrencileri birbirinden ayırt edilebilme ve öğrenciye güçlü ve zayıf olduğu noktalar hakkında geri bildirimler sunabilme özellikleri açısından SOLO taksonomisine dayalı rubrikleri standart rubriklere göre daha etkili bulmuşlardır. Hall ve Salmon'a (2003) göre, öğrencilere güçlü ve zayıf olduğu noktalar hakkında etkili geri bildirimler sunulması değerlendirme sürecinde rubrik kullanımının temel amaçlarından biridir. Puanlayıcı görüşlerine göre, SOLO taksonomisine dayalı rubriklerin standart rubriklere kıyasla bu amaca daha çok hizmet eden bir puanlama ölçeği olduğu söylenebilir.

BEŞİNCİ BÖLÜM

SONUÇ VE ÖNERİLER

Bu bölümde; araştırmadan elde edilen sonuçlar ile bu sonuçlar doğrultusunda geliştirilen önerilere yer verilmiştir. Ayrıca, bu araştırmanın kapsamı dâhilinde olmayan ancak ileri araştırmalarda ele alınabilecek önerilere değinilmiştir.

5.1. SONUÇLAR

Bu araştırmada, standart ve SOLO taksonomisine dayalı rubrikler ile puanlanan açık uçlu matematik sorularında puanlayıcı etkileri çok yüzeyli Rasch modeli ile incelenmiş ve bu iki rubrik türüne ilişkin puanlayıcı görüşleri belirlenmeye çalışılmıştır. Araştırmadan elde edilen sonuçlar, alt problemlerin sırasına uygun olarak aşağıda özetlenmiştir.

1. Standart rubrikler kullanılarak yapılan puanlamalarda;
- Öğrencilerin ve maddelerin yüksek güvenirlikte birbirinden ayırt edilebildiği,
- Puanlayıcılar arası güvenirliğin, mutlak uyumun ve kappa istatistiğinin düşük olduğu,
- Katılık ve cömertlikleri yönüyle puanlayıcılar arasında anlamlı fark bulunduğu belirlenmiştir.

2. Standart rubriklere göre yapılan puanlamalarda puanlayıcı katılığı/cömertliği ile merkeze yönelme etkisi, *puanlayıcı×madde* yanlılığının bulunduğu; halo etkisi, tutarsızlık, *puanlayıcı×birey* ve *puanlayıcı×birey×madde* yanlılıklarının ise söz konusu olmadığı saptanmıştır.

3. SOLO taksonomisine dayalı rubrikler kullanılarak yapılan puanlamalarda;
- Öğrencilerin ve maddelerin yüksek güvenirlikte birbirinden ayırt edilebildiği,
- Puanlayıcılar arası güvenirliğin, mutlak uyumun ve kappa istatistiğinin yüksek olduğu
- Puanlayıcıların benzer katılık ve cömertlikte puanlamalar yaptıkları tespit edilmiştir.

4. SOLO taksonomisine dayalı rubrikler kullanılarak yapılan puanlamalarda; puanlayıcı katılığı ve cömertliği, merkeze yönelme etkisi, tutarsızlık ve yanlılık şeklinde sıralanan puanlayıcı etkilerinden hiçbirinin bulunmadığı sonucuna ulaşılmıştır.
5. SOLO taksonomisine dayalı rubriklere göre elde edilen yetenek kestirimlerinin standart rubriklere göre hesaplanan yetenek kestirimlerine kıyasla öğrencilerin matematik başarılarını yordama gücünün daha yüksek olduğu belirlenmiştir. İki rubrik türüne göre hesaplanan yetenek kestirimleri arasında pozitif yönde ve güçlü bir korelasyon bulunmuştur.
6. SOLO taksonomisine dayalı rubrikler kullanılarak elde edilen yetenek kestirimleri ile standart rubriklere göre hesaplanan yetenek kestirimleri arasında istatistiksel olarak anlamlı fark bulunduğu tespit edilmiştir. Beşinci ve altıncı alt problemde ulaşılan sonuçlar bir arada ele alındığında, iki rubrik türüne göre hesaplanan yetenek kestirimleri arasında göreli uyum olduğu; ancak mutlak uyum bulunmadığı şeklinde bir çıkarım yapılmıştır.
7. Puanlayıcıların objektiflik, ayırt edicilik, etkili geri bildirim sunabilme, hazırlama ve kullanım kolaylığı gibi özellikler açısından SOLO taksonomisine dayalı rubrikleri standart rubriklere göre daha tercih edilebilir bir puanlama ölçeği olarak gördükleri belirlenmiştir.

5.2. ÖNERİLER

Araştırmadan elde edilen sonuçlar doğrultusunda, ölçme değerlendirme süreçlerine ve ileri araştırmalara ışık tutabilecek öneriler getirilmiştir. Öneriler; uygulamaya yönelik öneriler ve ileri araştırmalara yönelik öneriler olmak üzere iki başlıkta sunulmuştur.

5.2.1. Uygulamaya Yönelik Öneriler

Araştırmada elde edilen sonuçlar, uygulamaya yönelik bir takım önerileri beraberinde getirmektedir. Öncelikle, ÖSYM, 2013 yılı itibariyle, Açık Uçlu Sorularla Sınav Projesi'ni başlatmıştır. Bu proje kapsamında, ilerleyen yıllarda merkezi sınavlarda çoktan seçmeli soruların yanı sıra açık uçlu sorulara da yer verilmesi öngörülmektedir. Benzer şekilde, MEB ilerleyen yıllarda temel eğitimden ortaöğretime geçiş sistemi kapsamında uygulanan ortak sınavlarda açık uçlu sorulara yer vermeyi planlamaktadır. Araştırmada ulaşılan sonuçlar, merkezi sınavlarda açık

uçlu matematik sorularına yer verilmesi halinde, bu soruların puanlanmasında SOLO taksonomisine dayalı rubriklerden yararlanılabileceğini göstermiştir.

Ulusal düzeydeki sınavların yanı sıra, sınıf içerisinde yapılan ölçme-değerlendirme çalışmalarında da, SOLO taksonomisine dayalı rubrikler kullanılabilir. Böylelikle, öğrencilerin performanslarına ilişkin alınacak kararlar puanlayıcı etkisinden önemli ölçüde arındırılabilir. Yine sınıf içi değerlendirmelerde öğrencilerin yetenek düzeylerine ilişkin kestirimde bulunurken, açık uçlu soruların yalnızca dersin öğretmeni tarafından değil; zümre öğretmenler kurulundaki öğretmenlerin tümü tarafından puanlanması ve yapılan puanlamaların çok yüzeyli Rasch modeli ile analiz edilmesi önerilebilir. Bu sayede, puanlamayı yapan öğretmenlerin katılık ve cömertlikleri arasındaki farklar belirlenip, bu farklılıklara düzeltmeler uygulanarak daha adil değerlendirmeler yapılması mümkün olabilir.

Puanlayıcıların öğrencilere güçlü ve zayıf oldukları noktalar hakkında geri bildirim vermede standart rubrikleri SOLO taksonomisine dayalı rubriklere göre daha etkili bulduğu dikkate alındığında, biçimlendirme ve yetiştirmeye yönelik değerlendirmelerde daha çok SOLO taksonomisine dayalı rubriklerin kullanılması yararlı olacaktır. Son olarak, hangi rubrik türü kullanılırsa kullanılsın, puanlama işleminden önce kullanılacak rubrik türü hakkında puanlayıcılara eğitim verilmelidir. Eğitim kapsamında, puanlayıcılara örnek uygulamalar yaptırılmalı, örnek uygulamalar sonrasında, puanlayıcılara yaptıkları puanlamalar ile ilgili geri dönütler verilmelidir. Bu şekilde uygulanacak bir puanlayıcı eğitimi sayesinde, puanlama ölçeğinin kategorileri ve ölçülecek performansın boyutları hakkında puanlayıcılar arasında ortak bir algı oluşturulabilir.

Uygulamaya yönelik olarak getirilebilecek önerilerin bir kısmı açık uçlu soruların puanlanmasında kullanılacak rubrik türü ile ilgili iken; bir kısmı da puanlama işleminde nasıl bir yol takip edilmesi gerektiğiyle alakalıdır. Çünkü SOLO taksonomisine dayalı rubrikler kullanılarak yapılan puanlamalarda araştırma kapsamında incelenen puanlayıcı etkilerinin hiçbirine rastlanmaması ya da standart rubriklere göre yapılan puanlamalarda tutarsızlık ve halo etkisi ile *puanlayıcı×birey* ve *puanlayıcı×birey×madde* yanlılıklarının bulunmaması değerlendirme işleminde kullanılan rubriklerin yanı sıra puanlama işleminde takip edilen basamaklarla da ilişkili olabilir. Örneğin; açık uçlu sorulara verilen cevapların öğrenciden öğrenciye değil de; sorudan soruya puanlanması hem standart ve hem de SOLO taksonomisine dayalı rubrikler kullanılarak yapılan puanlamalara halo etkisinin karışmasını önleyen

önemli bir faktör olarak belirlenmiştir. Dolayısıyla; açık uçlu sorular puanlanırken, tüm öğrencilerin bir soruya verdikleri cevaplar puanladıktan sonra bir başka soruya geçilmesi halo etkisinin önlenebilmesi açısından yararlı bir uygulama olacaktır.

Araştırma sonuçlarına göre; puanlayıcıların öğrencilerin kimliği ve demografik özellikleri hakkında bilgi sahibi olmaması SOLO taksonomisine dayalı rubrikler ile yapılan puanlamalara yanlılık etkisinin karışmasını önleyen önemli bir faktördür. Benzer bir durum standart rubrikler kullanılarak yapılan puanlamalar için de geçerlidir. Puanlama işleminin, öğrencilerin cinsiyeti ve geçmiş matematik başarısı hakkındaki bilgiler sınav kâğıtlarından silindikten sonra gerçekleştirilmesi standart rubriklere göre yapılan puanlamalarda *puanlayıcı×birey* ile *puanlayıcı×birey×madde* yanlılığına rastlanmamasında etkili olan önemli bir faktör olarak tespit edilmiştir. Buna göre; değerlendirme işlemine yanlılık etkisinin karışması olasılığını azaltabilmek için açık uçlu soru içeren sınav kâğıtlarının isimsiz olarak puanlanması önerilebilir.

5.2.2. Araştırmanın Sınırlılıkları ve İleri Araştırmalara Yönelik Öneriler

Alanyazın incelendiğinde, açık uçlu matematik sorularının puanlanmasında SOLO taksonomisine dayalı rubriklerin kullanıldığı görülmektedir. Ancak bu çalışmalarda, SOLO taksonomisine dayalı rubriklerin puanlayıcı güvenirliğini nasıl etkilediğini belirlemeye yönelik herhangi bir işlem yapılmadığı belirlenmiştir. Özellikle, hangi rubrik türü kullanılırsa kullanılsın rubriklerin puanlayıcı güvenirliğini arttıracağı dikkate alındığında; puanlayıcı etkilerini azaltma konusunda SOLO taksonomisine dayalı rubriklerin standart rubriklerden farklı olarak ne gibi bir katkı sağladığı sorusuna literatürdeki çalışmalar yanıt olamamaktadır. Ayrıca, literatürde SOLO taksonomisine dayalı rubriklerin puanlayıcı güvenirliğine etkisinin incelendiği sınırlı sayıdaki araştırmada (Burnett, 1999; Chan vd., 2002; Leung, 2000; Yazıcı, 2013) puanlayıcılar arası korelasyon katsayısı ve basit uyum yüzdesi gibi klasik test kuramına dayalı tekniklerin kullanıldığı tespit edilmiştir. Buna bağlı olarak, literatürdeki araştırmalar SOLO taksonomisine dayalı rubriklerin puanlayıcı katılığı ve cömertliği, merkeze yönelme etkisi, halo etkisi, tutarsızlık ve yanlılık gibi puanlayıcı etkilerini azaltmada ne derece etkili olduğu konusuna ışık tutamamaktadır. Dolayısıyla, standart ve SOLO taksonomisine dayalı rubrikler kullanılarak puanlanan açık uçlu matematik sorularında puanlayıcı etkilerinin çok yüzeyli Rasch modeli ile incelendiği bu araştırmanın önemli olduğu

düşünülmektedir. Bununla birlikte, çalışmanın bir takım sınırlılıkları bulunmakta ve bu sınırlılıkların ileri araştırmalarla aşılabileceğine inanılmaktadır.

İlk olarak bu araştırmada, standart ve SOLO taksonomisine dayalı rubriklerin puanlayıcı etkilerini azaltma konusundaki işlevselliği açık uçlu matematik soruları üzerinden incelenmiştir. SOLO taksonomisini oluşturan düzeylerin cebirsel düşünme, istatistiksel düşünme ve geometrik düşünme gibi matematik dersi kapsamında yer alan farklı düşünme biçimleri ile paralellik gösterdiği bilinmektedir. Dolayısıyla SOLO taksonomisinin matematik dersi öğrenme çıktılarının değerlendirilmesine uygun yapısı, araştırma sonuçlarını etkileyen faktörlerden biri olabilir. Bu bağlamda; benzer çalışmaların farklı dersler için de yapılması, araştırmadan elde edilen bulguların genellenebilirliğine katkı sunması açısından oldukça önemlidir.

İkinci olarak bu araştırma, standart ve SOLO taksonomisine dayalı rubriklerin karşılaştırılması ile sınırlı tutulmuştur. Alanyazında, değerlendirme sürecinde rubrik kullanımının puanlayıcı etkilerini azalttığı ve puanlayıcı güvenirliğini artırdığı ifade edilmektedir (Moskal ve Leydens, 2000). Ancak, literatürde puanlayıcılar arası farklılıkları azaltmada herhangi bir taksonomi temele alınmadan hazırlanan standart rubriklerin mi; yoksa SOLO taksonomisi ve Bloom taksonomisi gibi taksonomilerden herhangi biri temele alınarak hazırlanan rubriklerin mi daha etkili olduğu sorusuna yanıt olabilecek bir çalışmaya rastlanmamıştır. Standart ve SOLO taksonomisine dayalı rubrikler ile puanlanan açık uçlu matematik sorularında puanlayıcı etkilerinin karşılaştırmalı olarak incelendiği bu çalışmanın literatürdeki söz konusu boşluğun giderilmesine katkı sağlayacak bir araştırma olduğu düşünülmektedir. Bununla birlikte, bu araştırma Bloom taksonomisine dayalı rubriklerin puanlayıcı etkilerini azaltmada standart rubriklere göre daha etkili olup olmadığı sorusunu yanıtlamada yetersiz kalmaktadır. Bu kapsamda ileri araştırmalarda, standart ve Bloom taksonomisine dayalı rubrikler kullanılarak puanlanan açık uçlu matematik sorularının puanlayıcı etkileri açısından karşılaştırılması önerilebilir. Yine ileri araştırmalarda, SOLO ve Bloom taksonomisine dayalı rubrikler ile puanlanan açık uçlu matematik soruları puanlayıcı etkileri açısından karşılaştırılabilir.

Üçüncü olarak, bu araştırmada SOLO taksonomisine dayalı rubrikler hazırlanırken Biggs ve Collis (1982) tarafından ileri sürülen beş düzeyli orijinal yapıya sadık kalınmıştır. Ancak, literatürde, SOLO taksonomisinin yedi, sekiz veya

dokuz düzeyleri olarak yeniden yapılandırıldığı çalışmalara da (Burnett, 1999; Chan vd., 2002) rastlanmaktadır. Literatürdeki bu çalışmalar SOLO taksonomisine dayalı rubriklerin hazırlanmasında dikkate alınan düzey sayısının değerlendirme sonuçlarını etkilediğini ortaya koymuştur. Bu nedenle, ileri araştırmalarda yedi, sekiz veya dokuz düzeyden oluşacak şekilde yapılandırılan SOLO taksonomisine dayalı rubrikler kullanılarak puanlanan açık uçlu matematik sorularında puanlayıcı etkilerinin incelenmesi yerinde olacaktır.

Çalışmanın, 104 öğrencinin açık uçlu sekiz matematik sorusuna verdikleri yanıtın yedi puanlayıcı tarafından puanlanması sonucu elde edilen verilerden oluşması araştırmaya ilişkin diğer bir sınırlılıktır. Rasch analizlerinde, 100 ile 200 öğrenciden elde edilen veriler parametre kestirimleri için yeterli görülmektedir. Bununla birlikte, madde tepki kuramına dayalı analizlerin katılımcı sayısının fazla olduğu örneklemlerde daha doğru kestirimler ürettiği (DeMars, 2010) ve çok yüzeyli Rasch modelinin madde tepki kuramının bir uzantısı olduğu dikkate alındığında, benzer bir çalışmanın daha büyük gruplarda yapılması önerilebilir.

Son olarak, bu araştırmada kullanılan standart ve SOLO taksonomisine dayalı rubrikler araştırmacı tarafından hazırlanmış olup araştırmaya dâhil edilen puanlayıcılar rubriklerin geliştirilmesinde görev almamıştır. Ancak, literatürde de ifade edildiği gibi, puanlayıcıların rubrik geliştirme sürecine dâhil edilmesi, puanlama ölçeğindeki kategorileri tam olarak anlamaları ve tutarlı bir biçimde kullanabilmeleri açısından kritik bir öneme sahiptir. Rubrik geliştirme sürecine puanlayıcıların dâhil edilmesi birçok değerlendirme sisteminde başvurulan yaygın bir uygulamadır (North, 2000). Bu çalışmada ise, puanlayıcıların rubrik geliştirme sürecinde görev almadığı bir yaklaşım benimsenmiştir. Benimsenen bu yaklaşım, değerlendirme sonuçlarını etkilemiş olabilir. Puanlayıcıların rubrik geliştirme sürecine dâhil edilmesi halinde, araştırmadan elde edilen sonuçların farklılık gösterebileceği düşünülmektedir. Puanlayıcıların standart ve SOLO taksonomisine dayalı rubriklerin geliştirilmesi sürecine dâhil edildiği ileri çalışmaların yapılmasıyla araştırmaya ilişkin bu sınırlılığın aşılabileceğine inanılmaktadır.

KAYNAKLAR

Abu Kassim, N.L. (2007, June). *Exploring rater judging behaviour using the many-facet Rasch model.* Paper Presented in the Second Biennial International Conference on Teaching and Learning of English in Asia: Exploring New Frontiers (TELiA2), Holiday Villa Beach & Spa Resort, Langkawi. Faculty of Communication and Modern Languages, Universiti Utara Malaysia.

Airasian, P. W. (2005). *Classroom assessment*. New York: McGraw-Hill.

Akbulut, Y. (2010). *Sosyal bilimlerde SPSS uygulamaları.* İstanbul: İdeal Kültür Yayıncılık.

Akın, Ö. ve Baştürk, R. (2012). Keman eğitiminde temel becerilerin Rasch ölçme modeli ile değerlendirilmesi. *Pamukkale Üniversitesi Eğitim Fakültesi Dergisi*, 31, 175-187.

Akkaş, E.N. (2009). 6.-*8. sınıf öğrencilerinin istatistiksel düşüncelerinin incelenmesi*. Yayınlanmamış Yüksek Lisans Tezi, Abant İzzet Baysal Üniversitesi, Sosyal Bilimler Enstitüsü, Bolu.

Alharby, E.R. (2006). *A comparison between two scoring methods, holistic vs. analytic using two measurement models, generalizability theory and the many facet Rasch measurement within the context of performance assessment*. Unpublished Doctoral Dissertation. Pennsylvenia: Pennsylvenia State University.

Alderson, C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge University Press, Cambridge.

Alkan, H. (1999). Matematikte ölçme ve değerlendirme, A.Özdaş (Ed.), *Matematik öğretimi* içinde (ss. 93-110). Eskişehir: Anadolu Üniversitesi Açık Öğretim Fakültesi Yayınları No:591.

Alsaadi, A. (2011). A comparison of primary mathematics curriculum in England and Qatar: The SOLO Taxonomy. *Proceedings of the British Society for Research into Learning Mathematics,* 21(3). 15 Temmuz 2014 tarihinde http://www.bsrlm.org.uk/IPs/ip21-3/BSRLM-IP-21-3-1.pdf adresinden alınmıştır.

Altun, M. (2005). *İlköğretim ikinci kademede matematik öğretimi.* Bursa: Alfa Basım Yayım.

Andrade, H., & Du, Y. (2005). Student perspectives on rubric-referenced assessment. *Practical Assessment, Research & Evaluation*, 3(5), 1-11.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43(9), 561-573.

Antonioni, D., & Woehr, D.J. (2001). Improving the quality of multisource rater performance. In D.W. Bracken, C.W. Timmreck and A.H. Church (Eds.), *The handbook of multisource feedback: The comprehensive resource for designing and ımplementing MSF processes* (pp. 114- 129). California: Jossey Bass Inc.

Araujo, E.A.C., Andrade, D.F., & Bortolotti, S.L.V. (2009). Item response theory. *Revista da Escola de Enfermagen da USP*, 43, 1000-1008.

Archbald, D.A., & Grant, T.J. (2000) What's on the test? An analytical framework and findings from an examination of teachers' math tests. *Educational Assessment*, 6(4), 221-256.

Arı, A. (2013). Bilişsel alan sınıflamasında yenilenmiş Bloom, SOLO, Fink, Dettmer taksonomileri ve uluslararası alanda tanınma durumları. *Uşak Üniversitesi Sosyal Bilimler Dergisi*, 6(2), 259-290.

Arias, R.M. (2010). Performance assessment. *Papeles del Psicólogo*, 31(1), 85-96.

Arter, J., & McTighe, J. (2001). *Scoring rubrics in the classroom: Using performance criteria for assessing and improving student performance*. Thousand Oaks, CA: Corwin Press, Inc.

Aşan, E. (2014). *İlköğretim vatandaşlık ve demokrasi eğitimi dersi 8. sınıf ders kitabı*. Ankara: Ekoyay Eğitim Yayıncılık, Matbaacılık.

Atılgan, H. (2004). *Genellenebilirlik kuramı ve çok değişkenlik kaynaklı Rasch modelinin karşılaştırılmasına ilişkin bir araştırma*. Yayınlanmamış Doktora Tezi, Hacettepe Üniversitesi, Sosyal Bilimler Enstitüsü, Ankara.

Atılgan, H. (2005a). Genellenebilirlik kuramı ve puanlayıcılar arası güvenirlik için örnek bir uygulama. *Eğitim Bilimleri ve Uygulama*, 4(7), 95-108.

Atılgan, H. (2005b). Müzik öğretmenliği özel yetenek seçme sınavının çok yüzeyli Rasch modeli ile analizi (İnönü Üniversitesi örneği). *Eğitim Araştırmaları Dergisi*, 20, 62-73.

Atılgan, H. (2009). Test geliştirme. H. Atılgan (Ed.), *Eğitimde ölçme ve değerlendirme* içinde (ss. 315-348). Ankara: Anı Yayıncılık.

Atmaz, G. (2009). *Puanlama yönergesi (rubrik) kullanılması durumunda puanlayıcı güvenirliğinin incelenmesi*. Yayınlanmamış Yüksek Lisans Tezi, Mersin Üniversitesi, Mersin, Türkiye.

Bacha, N. (2001). Writing evaluation: What can analytic versus holistic essay scoring tell us? *System*, 29(3), 371-383.

Bachman, L.F. (2004). *Statistical analyses for language assessment*. Cambridge, England: Cambridge University Press.

Bağcan Büyükturan, E. ve Çıkrıkçı Demirtaşlı, N. (2013). Çoktan seçmeli testler ile yapılandırılmış gridlerin psikometrik özellikleri bakımından karsılaştırılması. *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi*, 46(1), 395-415.

Bağdat, O. (2013). *İlköğretim 8. sınıf öğrencilerinin cebirsel düşünme becerilerinin SOLO taksonomisi ile incelenmesi*. Yayınlanmamış Yüksek Lisans Tezi, Eskişehir Osmangazi Üniversitesi Eğitim Bilimleri Enstitüsü, Eskişehir.

Baird, J.A., Hayes, M., Johnson, R., Johnson, S., & Lamprianou, I. (2013). *Marker effects and examination reliability a comparative exploration from the perspectives of generalizability theory, Rasch modelling and multilevel modelling.* 3 Mayıs 2013 tarihinde http://www.ofqual.gov.uk/files/2013-01-21-marker-effects-and-examination-reliability.pdf adresinden alınmıştır.

Baker, F.B. (2001). *The basics of item response theory*. ERIC Clearinghouse on Assessment and Evaluation, University of Maryland, College Park, MD.

Banerjee, M., Capozzoli, M., McSweeney, L., & Sinha, D. (1999). Beyond kappa: A review of interrater agreement measures. *The Canadian Journal of Statistics*, 27(1), 3-23.

Bargainnier, S. (2003). *Fundamentals of rubrics*. 22 Ağustos 2013 tarihinde http://www.webpages.uidaho.edu/ele/scholars/Practices/Evaluating_Projects/Resources/Using_Rubrics.pdf adresinden alınmıştır.

Barkaoui, K. (2007). Participants, texts, and processes in ESL/EFL essay tests: A narrative review of the literature. *The Canadian Modern Language Review/La Revue canadienne des langues vivantes*, 64(1), 99-134.

Barkaoui, K. (2008). *Effects of scoring method and rater experience on ESL essay rating processes and outcomes*. Unpublished Doctoral Dissertation, University of Toronto, Canada.

Barkaoui, K. (2010). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly*, 7(1), 54-74.

Baştürk, R. (2010). Bilimsel araştırma ödevlerinin çok yüzeyli Rasch ölçme modeli ile değerlendirilmesi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, 1(1), 51-57.

Baykul, Y. (2000). *Eğitimde ve psikolojide ölçme: Klasik test teorisi ve uygulaması*. Ankara: ÖSYM Yayınları.

Baykul, Y. ve Güzeller, C.O.(2013). *Sosyal bilimler için istatistik*. Ankara: Pegem Akademi Yayıncılık.

Bayram, N. (2009). *Sosyal bilimlerde SPSS ile veri analizi*. Bursa: Ezgi Kitabevi.

Bechger, T.M., Maris, G., & Hsiao, Y.P. (2010). Detecting halo effects in performance-based examinations. *Applied Psychological Measurement*, 34(8), 607-619.

Becker, A. (2010/2011). Examining rubrics used to measure writing performance in U.S. intensive english programs. *The CATESOL Journal*, 22(1), 113-130.

Bernardin, H.J., & Buckley, M.R. (1981). Strategies in rater traning. *Academy of Management Review*, 6(2), 205-212.

Bernardin, H.J., & Walter, C.S. (1977). Effects of rater training and diary-keeping on psychometric error in ratings. *Journal of Applied Psychology*, 62(1), 64-69.

Bhattacharyya, T., Bhattacharya, B., & Mitra, T. (2012, July). *Impact of SOLO taxonomy in computer aided instruction to qualitative outcome of learning for secondary school children*. Paper Presented at the 4th IEEE International Conference on Technology for Education. Hyderbad, India.

Biggs, J.B. (1979). Individual differences in study processes and the quality of learning outcomes. *Higher Education*, 8(4), 381-394.

Biggs, J.B. (1995). Assumptions underlying new approaches to educational assessment: Implications for Hong Kong. *Curriculum Forum*, 4(2), 1-22.

Biggs, J.B., & Collis, K.F. (1982). *Evaluating the quality of learning: The SOLO taxonomy*. Academic Press.

Biggs, J.B. (2003). *Teaching for quality learning at University*. Maidenhead: Open University Press.

Biggs, J., & Tang, C. (2007). *Teaching for quality learning at university*. New York: Society for Research into Higher Education & Open University Press.

Bingölbali, E., Özmantar, M.F. ve Akkoç, H. (2008). *Sınıf öğretmenlerinin farklı matematiksel çözüm yollarını değerlendirme süreçleri*. VII. Ulusal Sınıf Öğretmenliği Sempozyumu'nda sunulmuş sözlü bildiri, Çanakkale, Türkiye.

Birenbaum, M., & Feldman, R.A. (1998) Relationships between learning patterns and attitudes towards two assessment formats. *Educational Research*, 40(1), 90-98.

Birkimer, J.C., & Brown, J.H. (1979). Back to basics: Percentage agreement measures are adequate, but there are easier ways. *Journal of Applied Behavior Analysis*, 12(4), 535-543.

Bond, T.G., & Fox, C.M. (2001). A*pplying the Rasch model: Fundamental measurement in the human sciences*. London: Lawrence Erlbaum Associates.

Borlotti, S.L.V., Tezza, R., Andrade, D.F., Bornia, A.C., & Junior, A.F.S. (2013). Relevance and advantages of using the item response theory. *Quality & Quantity*, 47(4), 2341-2360.

Brabrand, C. (2008). Constructive alignment for teaching model-based design for concurrency. In K. Jensen, W. van der Aalst, and J. Billinton (Eds), *Transactions on petri nets and other models of concurrency* (pp. 1-18). Berlin, Heidelberg: Springer-Verlag.

Brabrand, C., & Dahl, B. (2009). Using the SOLO taxonomy to analyze competence progression of university science curricula. *Higher Education*, 58(4), 531-549.

Braun, H.I. (1988). Understanding scoring reliability: Experiments in calibrating essay readers. *Journal of Educational Statistics*, 13, 1-18.

Brennan, R.L. (2000). (Mis)conceptions about generalizability theory. *Educational Measurement: Issues and Practice*, 19(1), 5-10.

Brennan, R.L., Gao, X., & Colton, D.A. (1995). Generalizability analyses of work key listening and writing tests. *Educational and Psychological Measurement*, 55(2), 157-176.

Brentari, E., & Golia, S. (2008). Measuring job satisfaction in the social services sector with the Rasch model. *Journal of Applied Measurement*, 9(1), 45-56.

Brookhart, S. M. (1999). *The art and science of classroom assessment: The missing part of pedagogy*. ASHE-ERIC Higher Education Report, 27(1). Washington, DC: The George Washington University, Graduate School of Education and Human Development.

Brown, G.T.L., & Hirschfeld, G.H.F. (2008) Students' conceptions of assessment: Links to outcomes. *Assessment in Education: Principles, Policy & Practice*, 15(1), 3-17.

Brown, J.D., & Hudson, T. (2002). *Criterion-referenced language testing*. Cambridge, England: Cambridge University Press.

Brualdi Timmins, A.C. (1998). Implementing performance assessment in the classroom. *Practical Assessment, Research & Evaluation*, 6(2). 27 Ağustos 2013 tarihinde http://pareonline.net/getvn.asp?v=6&n=2 adresinden alınmıştır.

Burnett, P.C. (1999). Assessing the structure of learning outcomes from counselling using the SOLO taxonomy: An exploratory study. *British Journal of Guidance & Counselling*, 27(4), 567-580.

Burns, M. (1995). *Writing in math class*. Sausalito, CA: Math Solutions Publications.

Busching, B. (1998). Grading inquiry projects. *New Directions for Teaching and Learning*, 74, 89-96.

Büyüköztürk, Ş. (2010). *Sosyal bilimler için veri analizi el kitabı*. Ankara: Pegem Akademi Yayınları.

Büyüköztürk, Ş., Çakmak, E.K., Akgün, Ö.E., Karadeniz, Ş. ve Demirel, F. (2010). *Bilimsel araştırma yöntemleri*. Ankara: Pegem Akademi Yayınları.

Cagnone, S., & Ricci, R. (2005). Student ability assessment based on two IRT models. *Metodoloˇski Zvezki*, 2(2), 209-218.

Cardinet, J., Johnson, S., & Pini, G. (2010). *Applying generalizability theory using EduG*. New York, NY: Taylor and Francis.

Carr, N. (2000). A comparison of the effects of analytic and holistic rating scale types in the context of composition tests. *Issues in Applied Linguistics*, 11(2), 207-241.

Cellar, D.F., Curtis, J.R., Kohlepp, K., Poczapski, P., & Mohiuddin, S. (1989). The effects of rater training, job analysis format and congruence of training on job evaluation ratings. *Journal of Business & Psychology*, 3(4), 387-401.

Chamoso, J.M., & Caceres, M.J. (2009). Analysis of the reflections of student-teachers of mathematics when working with learning portfolios in Spanish University classrooms. *Teaching and Teacher Education*, 25(1), 198-206.

Chan, A.L.F., Chien, T.W., Su., C.Y., & Lin, S.J. (2009). Validation of a shortened Taiwanese version of an asthma quality-of-life questionnaire by Rasch model. *Allergy and Asthma Proceedings*, 30(2), 171-180.

Chan, C.C., Hong, J.H., & Chan, M.Y.C. (2001) *Applying the structure of the observed learning outcomes (SOLO) taxonomy on student's learning outcomes: A comparative review*. Unpublished manuscript, Hong Kong, Hong Kong Polytechnic University.

Chan, C.C., Tsui, M.S., Mandy, Y.C., & Hong, J.H. (2002). Applying the structure of the observed learning outcomes (SOLO) taxonomy on student's learning outcomes: An empirical study. *Assessment and Evaluation in Higher Education*, 27(6), 511-527.

Chernyshenko, O.S., Stark, S., & Chan, K.Y. (2001). Fitting item response theory models to two personality inventories: Issues and insights. *Multivariate Behavioral Research*, 36(4), 523-562.

Chi, E. (2001). Comparing holistic and analytic scoring for performance assessment with many-facet Rasch model. *Journal of Applied Measurement*, 2(4), 379-388.

Clark-Carter, D. (2010). *Quantitative psychological research: The complete student's companion*. Hove: Psychology Press.

Cohen, A.D. (1994). *Assessing language ability in the classroom*. Boston: Heinle & Heinle.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin* 70(4), 213-220.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37-46.

Collis, K.F., & Romberg, T.A. (1992). *Collis-Romberg mathematical problem solving profiles*. Melbourne: Australian Council for Educational Research.

Congdon, P., & McQueen, J. (2000). The stability of rater severity in large-scale assessment programs. *Journal of Educational Measurement*, 37(2), 163-178.

Courtney, T.D. (1986). The significance of the SOLO taxonomy for learning and teaching in geography. *Geographical Education*, 5(2), 47-50.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Belmont, CA: Wadsworth.

Cronbach, L.I. (1990). *Essentials of psychological testing*. New York: Harper and Row.

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing*, 7(1), 31-51.

Çelik, D. (2007). *Öğretmen adaylarının cebirsel düşünme becerilerinin analitik incelenmesi*. Yayınlanmamış Doktora Tezi, Karadeniz Teknik Üniversitesi, Fen Bilimleri Enstitüsü, Trabzon.

Çetin, B. (2005). *Geleneksel yöntemle ve eleme yöntemi ile puanlanan çoktan seçmeli testlerin psikometrik özelliklerinin incelenmesi*. Yayınlanmamış Doktora Tezi, Hacettepe Üniversitesi, Sosyal Bilimler Enstitüsü, Ankara.

Danielson, C., & Hansen, P. (1999). *A collection of performance tasks and rubrics: Primary school mathematics*. Larchmont, NY: Eye on Education.

David, A.B. (2008). Comparison of classification accuracy using Cohen's weighted kappa. *Expert Systems with Applications*, 34, 825-832.

DeMars, C. (2010). *Item response theory*. Oxford, UK: Oxford University Press.

DeVellis, R.F. (2003). *Scale development: Theory and applications*. Newbury Park: Sage Publications.

Doğan, N. (2009). Çoktan seçmeli testler. H. Atılgan (Ed.), *Eğitimde ölçme ve değerlendirme* içinde (ss. 223-268). Ankara: Anı Yayıncılık.

Domino, G., & Domino, M.L. (2006). *Psychological testing: An introduction*. Cambridge: Cambridge University Press.

Dunbar, N.E., Brooks, C.F., & Miller, T.K. (2006). Oral communication skills in higher education: Using a performance-based evaluation rubric to assess communication skills. *Innovative Higher Education*, 31(2), 2006, 115-128.

Ebel, R.L. (1951). Estimation of the reliability of ratings. *Psychometrika*, 16(4), 407-424.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly*, 2(3), 197-221.

Eckes, T. (2009). Many-facet Rasch measurement. In S. Takala (Ed.), *Reference supplement to the manual for relating language examinations to the common european framework of reference for languages: Learning, teaching, assessment* (Section H). Strasbourg, France: Council of Europe/Language Policy Division.

Elder, C., Barkhuizen, G., Knoch, U., & von Randow, J. (2007). Evaluating rater responses to an online training program for L2 writing assessment. *Language Testing*, 24(1), 37-64.

Elder, C., Knoch, U., Barkhuizen, G., & von Randow, J. (2005). Individual feedback to enhance rater training: Does it work?, *Language Assessment Quarterly*, 2(3), 175-196.

Embretson, S.E., & Reise, S.P. (2000). *Item response theory for psychologists*. Mahwah, NJ:Lawrence Erlbaum Associates, Inc.

Engelhard, G. (1992). The measurement of writing ability with a many-facet Rasch model. *Applied Measurement in Education*, 5(3), 171-191.

Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement*, 31(2), 93-112.

Engelhard, G. (2002). Monitoring raters in performance assessments. In G. Tindal & T.M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 261-287). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Engelhard, G. (2005). Item response theory models (IRT) for rating scale data. *Encyclopedia of Statistics in Behavioral Science*, Vol., 2 (pp. 995-1003). Chichester: John Wiley & Sons, Ltd.

Engelhard, G., & Stone, G.E. (1998). Evaluating the quality of ratings obtained from standard-setting judges. *Educational and Psychological Measurement*, 58(2), 179-196.

Engelhard, G. (2011). Evaluating the bookmark judgments of standard-setting panelists. *Educational and Psychological Measurement*, 71(6), 909-924.

Fahim, M., & Bijani, H. (2011). The effects of rater training on raters' severity and bias in second language writing assessment. *Iranian Journal of Language Testing*, 1(1), 1-16.

Farrokhi, F., & Esfandiari, R. (2011). A many-facet Rasch model to detect halo effect in three types of raters. *Theory and Practice in Language Studies*, 1(11), 1531-1540.

Farrokhi, F., Esfandiari, R., & Vaez Dalili, M. (2011). Applying the many-facet Rasch model to detect centrality in self-assessment, peer-assessment and teacher assessment. *World Applied Sciences Journal*, 15, 70-77.

Fisicaro, S.A., & Lance, C.E. (1990). Implications of three causal models for the measurement of halo error. *Applied Psychological Measurement*, 14(4), 419-429.

Fleiss, J.L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5), 378-382.

Fox, J.P. (2010). *Bayesian item response modeling*. New York: Springer.

Fraser, L., Harich, K., Norby, J., Brzovic, K., Rizkallah, T., & Loewy, D. (2005). Diagnostic and value-added assessment of business writing. *Business Communication Quarterly*, 68(3), 290-305.

Galati, D.E. (2007). *A G-theory based adjustment process for calibrating marker scores.* Unpublished Doctoral Dissertation. University o f Toronto, Ontario, Canada.

Gezer, M. ve İlhan, M. (2014). 8. sınıf vatandaşlık ve demokrasi eğitimi dersi kazanımları ile değerlendirme sorularının SOLO taksonomisine göre incelenmesi. *Doğu Coğrafya Dergisi*, 19 (32), 193-207.

Gezer, M., İlhan, M., Öner Sünkür, M. ve Çetin, B. (2014). *Sosyal bilgiler dersi sınav sorularının SOLO taksonomisine göre incelenmesi*. III. Ulusal Eğitim Programları ve Öğretim Kongresi'nde sunulmuş sözlü bildiri, Gaziantep Eğitim Fakültesi, Gaziantep, Türkiye.

Gilber, R. (2004). Teaching thinking in SOE. In Rob Gilbert (Eds.), *Studying society and environment: A guide for teachers* (pp. 56-79). Southbank: Thomson/Social Science Press.

Goel, S. (2011). *An overview of selected theories about student learning Indo-US workshop on effective teaching and learning at college/university level, IIIT Delhi,* http://sites.iiitd.ac.in/indo-us/papers/Paper_Sanjay%20Goel.pdf adresinden 5 Ekim 2013 tarihinde alınmıştır.

Goodwin, L.D. (2001). Interrater agreement and reliability. *Measurement in Physical Education and Exercise Science*, 5(1), 13-34.

Goulden, N.R. (1992). Theory and vocabulary for communication assessments. *Communication Education*, 41(3), 258-269.

Goulden, N.R. (1994). Relationship of analytic and holistic methods to raters' scores for speeches. *The Journal of Research and Development in Education*, 27(2), 73-82.

Graham, M., Milanowski, A., & Miller, J. (2012). *Measuring and promoting inter-rater agreement of teacher and principal performance ratings*. Washington, DC: Center for Educator Compensation Reform. 10 Haziran 2013 tarihinde http://cecr.ed.gov/pdfs/Inter_Rater.pdf adresinden alınmıştır.

Groth, R.E., & Bergner, J.A. (2006). Preservice elementary teachers' conceptual and procedural knowledge of mean, median and mode. *Mathematical Thinking and Learning*, 8(1), 37-63.

Gunning, T. G. (2001). *Assessing and correcting reading and writing difficulties*. Boston: Allyn & Bacon.

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

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

Güler, N. ve Gelbal, S. (2010). Klasik test kuramı ve çok değişkenlik kaynaklı Rasch modeli üzerine bir çalışma. *Egitim Araştırmaları-Eurasian Journal of Educational Research,* 38, 108-125.

Güler, N. (2012). *Eğitimde ölçme ve değerlendirme*. Ankara: Pegem Akademi Yayınları.

Güler, N. (2014). Analysis of open-ended statistics questions with many facet Rasch model. *Eurasian Journal of Educational Research,* 55, 73-90.

Gültekin, S. (2012). Testlerde kullanılacak madde türleri, hazırlama ilkeleri ve puanlanması. N. Çıkrıkçı Demirtaşlı (Ed.), *Eğitimde ölçme ve değerlendirme* içinde (ss. 169-245). Ankara: Elhan Yayınları.

Hafner, J., & Hafner, P. (2003). Quantitative analysis of the rubric as an assessment tool: An empirical study of student peer-group rating. *International Journal of Science Education,* 25(12), 1509-1528.

Haiyang, S. (2010). An application of classical test theory and many facet Rasch measurement in analyzing the reliability of an English test for non-English major graduates. *Chinese Journal of Applied Linguistics*, 33(2), 87-102.

Hall, E.K., & Salmon, S.J. (2003). Chocolate chip cookies and rubrics helping students understand rubrics in inclusive settings. *Teaching Exceptional Children, 35*(4), 8-11.

Hallgren, K.A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology,* 81, 23-34.

Halonen, J.S., Bosack, T., Clay, S., McCarthy, M., Dunn, D.S., Hill IV, G.W., vd. (2003). A rubric for learning, teaching, and assessing scientific ınquiry in psychology. *Teaching of Psychology*, 30(3), 196-208.

Hambleton, R.K. (1987). The three parameter logistic model. In D.L. McArthur (Eds.), *Alternative approaches to the assessment of achievement* (pp. 129-158). Boston: Kluwer Academic Publishers.

Hambleton, R.K. (1995). Meeting the measurement challenges of the 1990s and beyond: New assessment models and methods. In T. Oakland & R.K. Hambleton (Eds.), *International perspectives on academic assessment* (pp. 83-104). Boston, MA: Kluwer Academic Publishers.

Hambleton, R.K., & Jones, R.W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, 12(3), 38-47.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: SAGE Publications, Inc.

Hamilton, J., Reddel, S., & Spratt, M. (2001). Teachers' perceptions of on-line rater training and monitoring. *System*, 29(4), 505-520.

Hamp-Lyons, L., & Zhang, B.W. (2001). World Englishes: Issues in and from academic writing assessment. In J. Flowerdew & M. Peacock (Eds.), *Research perspectives on English for academic purposes* (pp. 101-116). Cambridge, England: Cambridge University Press.

Harik, P. (2008). *Statistical modeling of rater stringency in performance assessments*. Unpublished Doctoral Dissertation. University of Delaware, Delaware, ABD.

Harvey, R.J., & Hammer, A.L. (1999). Item response theory. *The Counseling Psychologist*. 27(3), 353-383.

Hattie, J. A. & Purdie, N. (1994). *Using the SOLO taxonomy to classify test items*. Unpublished manuscript, University of Western Australia, Graduate School of Education, Perth, Aus.

Hattie, J.A. & Purdie, N. (1998). The SOLO method and item construction. In G. Boulton-Lewis & B. Dart (Eds.), *Learning in Higher Education*. Hawthorn, Australia: ACER.

Hawkins, W., & Hedberg, J.G. (1986). Evaluating LOGO: Use of the SOLO taxonomy. *Australian Journal of Educational Technology (AJET)*, 2(2), 103-109.

Holbrook, J.B. (1989). *Writing chemistry items using the SOLO taxonomy*. A Symposium Presented to the Sixth Annual Conference of the Hong Kong Educational Research Association. Hong Kong: City Polytechnic of Hong Kong.

Hoyt, W.T. (2000). Rater bias in psychological research: When is it a problem and what can we do about it? *Psychological Methods*, 5(1), 64-86.

Howell, D.C. (2010). *Statistical methods for psychology*. Belmont, CA: Wadsworth, Cengage Learning.

Howell, R.J. (2011). Exploring the impact of grading rubrics on academic performance: Findings from a quasi-experimental, pre-post evaluation. *Journal on Excellence in College Teaching*, 22(2), 31-49.

Huang, J. (2009). Factors affecting the assessment of ESL students' writing. *International. Journal of Applied Educational Studies*, 5(1), 1-17.

Hundzynski, C. (2008). *Elementary teachers in a science inquiry study group: Concerns, uses, and reflections*. Unpublieshed Doctoral Dissertation. Fordham University, New York, ABD.

Hung, S.P., Chen, P.S., & Chen, H.C. (2012). Improving creativity performance assessment: A rater effect examination with many facet Rasch model. *Creativity Research Journal*, 24(4), 345-357.

Hux, K., Sanger, D., Reid, R., & Maschka, A. (1997). Discourse analysis procedures. Reliability issues. *Journal of Communication Disorders*, 30(2), 133-150.

Ireson, J. (2008). *Learners, learning and educational activity*. New York: Routledge.

İlhan, M. ve Çetin, B. (2014a, Eylül). *Performans değerlendirmeye karışan puanlayıcı etkileri: Kuramsal bir analiz*. YILDIZ International Conference on Educational Research and Social Studies, Yıldız Teknik Üniversitesi, İstanbul, Türkiye.

İlhan, M. ve Çetin, B. (2014b). Performans değerlendirmeye karışan puanlayıcı etkilerini azaltmanın yollarından biri olarak puanlayıcı eğitimleri. *Journal of Europan Education*, 4(2), 29-38.

Jackson, S.E., Schuler , R.S., & Werner, S. (2009). *Managing human resources*. Mason, OH: Cengage/Southwestern Publishers.

Jaskari, M.M. (2013). The challenge of assessing creative problem solving in client-based marketing development projects: A SOLO taxonomy approach. *Journal of Marketing Education*, 20(10) 1-14.

Jimoyiannis, A. (2011). Using SOLO taxonomy to explore students' mental models of the programming variable and the assignment statement. *Themes in Science & Technology Education*, 4(2), 53-74.

Johnson, J.S., & Lim, G.S. (2009). The influence of rater language background on writing performance assessment. *Language Testing*, 26(4), 485-505.

Johnson, R.L., Penny, J.A., & Gordon, B. (2009). *Assessing performance: Designing, scoring, and validating performance tasks*. New York: The Guilford Press.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, 2(2), 130-144.

Joy, D., & Wilmot, B. (2008). *Assessing progress toward college readiness with psychometric and cognitive models of student learning in mathematics*. Berkeley: University of California Publications.

Jurdak, M. (1991) Van Hiele levels and the SOLO taxonomy. *International Journal of Mathematical Education in Science and Technology*, 22(1), 57-60.

Kaliski, P.K., Wind, S.A,, Engelhard, G., Jr., & Morgan, D.L., Plake, B.S., & Reshetar, R.A. (2012). Using the many-faceted Rasch model to evaluate standard setting judgments: An illustrationwith the advanced placement environmental science exam. *Educational and Psychological Measurement,* XX(X), 1-26.

Kan, A. (2007). Performans değerlendirme sürecine katkıları açısından yeni program anlayışı içerisinde kullanılabilecek bir değerlendirme yaklaşımı: Rubrik puanlama yönergeleri. *Kuram ve Uygulamada Eğitim Bilimleri*, 7(1), 129-152.

Kan, A. (2009). Ölçme araçlarında bulunması gereken nitelikler. H. Atılgan (Ed.), *Eğitimde ölçme ve değerlendirme* içinde. Ankara: Anı Yayıncılık.

Kantrov, I. (2000). Assessing students' mathematics learning. *Issues in Mathematics Education* (pp. 1-11). Boston: Educational Development Center.

Kanuka, H. (2005). An exploration into facilitating higher levels of learning in a text-based internet learning environment using diverse ınstructional strategies. *Journal of Computer-Mediated Communication*, 10(3). 09 Ekim 2013 tarihinde http://onlinelibrary.wiley.com/doi/10.1111/j.1083-6101.2005.tb00256.x/full adresinden alınmıştır.

Kanuka, H. (2011). Interaction and the online distance classroom: Do instructional methods effect the quality of interaction? *Journal of Computing in Higher Education*, 23(2-3), 143-156.

Kaptan, S. (1998). *Bilimsel araştırma ve istatistik teknikleri*. Ankara: Tekışık Web Ofset Tesisleri.

Karakuş, M.A. (2008). *İlköğretim öğrencilerinin başarılarının değerlendirilmesinde hale etkisi*. Yayınlanmamış Yüksek Lisans Tezi, Anaka Üniversitesi, Ankara, Türkiye.

Karami, H. (2012). The relative impact of persons, items, subtests, and academic background on performance on a language proficiency test. *Psychological Test and Assessment Modeling*, 54(3), 211-226.

Kasap, Y. (2008). *Dereceli puanlama anahtarı ve puanlama anahtarından elde edilen puanların karşılaştırılması*. Yayınlanmamış Yüksek Lisans Tezi, Hacettepe Üniversitesi, Ankara, Türkiye.

Kayani, M.A.H., Ajmal, M., & Rahman, F. (2010). Teachers' perception regarding examination based on SOLO taxonomy. *International Journal of Academic Research*, 2(6), 208-211.

Khaatri, N., & Kane, M.B., & Reeve, A.L. (1995). How performance assessments affect teaching and learning. *Educational Leadership*, 53(3), 80-83.

Klein , S.P., Stecher, B.M., Shavelson, R.J., McCaffrey, D., Ormseth, T., Bell, R.M., Comfort, K., & Othman, A.R. (1998) Analytic versus holistic scoring of science performance tasks. *Applied Measurement in Education*, 11(2), 121-137.

Kiani, M.A.H. (2011). *A study to evaluate the examination system at grade-V in the Punjab based on SOLO taxonomy*. Unpublished Doctoral Dissertation, Fonudation University Islamabad, Pakistan.

Killen, R. (2009). *Effective teaching strategies: Lessons from research and practice*. South Melbourne: Cengage Learning Australia.

Killen, R., & Hatting, S.A. (2004). A theoretical framework for measuring the quality of student learning in outcomes-based education. *South African Journal of Higher Education*, 18(1), 72-86.

Kim, Y.H. (2009). An investigation into native and non-native teachers' judgments of oral English performance: A mixed methods approach. *Language Testing*, 26(2), 187-217.

Kim, Y., Park, I., & Kang, M. (2012). Examining rater effects of the TGMD-2 on children with ıntellectual disability. *Adapted Physical Activity Quarterly*, 29, 346-365.

Kind, P.M. (1999). Performance assessment in science-What are we measuring? *Studies in Educational Evaluation*, 25(3), 179-194.

Kline, T. (2005). *Psychological testing: A practical approach to design and evaluation*. Thousand Oaks: Sage Publications.

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing*, 26(20), 275-304.

Knoch, U., Read, J., & von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing*, 12(1), 26-43.

Kobayashi, T. (1992). Native and nonnative reactions to ESL compositions. *TESOL Quarterly*, 26(1), 81-112.

Koons, H.H. (2008). *The reading-writing connection: An investigation of the relationshıp between reading ability and writing quality across multiple grades and three writing discourse modes*. Unpublieshed Doctoral Dissertation. University of North Carolina, ABD.

Koskey, K.L.K. (2009). *A mixed-model approach to examining the use of absolute magnitude estimation scaling using the many-facet Rasch model and an ınstrumental case study*. Unpublieshed Doctoral Dissertation. The University of Toledo, Toledo, İspanya.

Kozaki, Y. (2004). Using GENOVA and FACETS to set multiple standards on performance assessment for certification in medical translation from Japanese into English. *Language Testing*, 21(1), 1-27.

Kozlowski, S.W.J. (2012). *The Oxford handbook of organizational psychology*. New York: Oxford University Press.

Kramer, J., Bowyer, P., Kielhofner, G., O'Brien, J., & Barbosa, V.B. (2009). Examining rater behavior on a revised version of the short child occupational profile (SCOPE). *OTJR: Occupation, Participation and Health,* 29(2), 88-96.

Kraska-Miller, M. (2014). *Nonparametric statistics for social and behavioral sciences.* Boca Raton, FL: Taylor & Francis.

Kutlu, Ö. (2006). Üst düzey zihinsel süreçleri belirleme yolları: Yeni durum belirleme yaklaşımları. *Çağdaş Eğitim*, 335, 15-21.

Kutlu, Ö., Doğan, C.H. ve Karakaya, İ. (2010). *Öğrenci başarısının belirlenmesi performansa ve portfolyoya dayalı durum belirleme*. Ankara: Pegem Akademi Yayınları.

Kumar, DSP D. (2005). Performance appraisal: The importance of rater training. *Journal of the Kuala Lumpur Royal Malaysia Police College*, 4, 1-15.

Labig, C. E. Jr., & Chye, T.Y. (1996). Problems with performance appraisal? Remedies for HR executives. *Research and Practice in Human Resource Management,* 4(1), 107-113.

Lake, D. (1999) Helping Students to go SOLO: Teaching critical numeracy in the biological sciences. *Journal of Biological Education*, 33(4), 191-198.

Landy, R.J., & Farr, J.L. (1980). Performance rating. *Psychological Bullutein* 87(1), 72-107.

LeBreton, J.M., & Senter, J.L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods*, 11(4), 815-852.

Leung, C.F. (2000). Assessment for learning: Using SOLO taxonomy to measure design performance of design & technology students. I*nternational Journal of Technology and Design Education*, 10(2), 149-161.

Lian, L.H., & Idris, N. (2006). Assessing algebraic solving ability of form four students. *International Electronic Journal of Mathematics Education*, 1(1), 55-76.

Lian, L.H., Meng, C.C., Yew, W,T., & Idris, N. (2009). *Assessing a hierarchy of pre-service teachers' algebraic thinking of equation*. 26 Kasım 2013 tarihinde http://www.recsam.edu.my/cosmed/cosmed09/abstractsfullpapers2009/abstract/Mathematics%20Parallel%20PDF/Full%20Paper/05.pdf adresinden alınmıştır.

Lian, L.H., & Yew, W.T. (2012). Assessing algebraic solving ability: A theoretical framework. *International Education Studies*, 5(6), 177-188.

Lian, L.H., Yew, W.T., & Idris, N. (2009). Kebolehan penyelesaian persamaan linear: Satu kerangka dalam penaksiran bilik darjah. *Malaysian Journal of Learning & Instruction (MJLI)*, 6, 79-101.

Lievens, F. (2001). Assessor training strategies and their effects on accuracy, interrater reliability and discriminant validity. *Journal of Applied Psychology*, 86(2), 255-264.

Linacre, J.M. (1989). *Many-facet Rasch measurement*. Chicago: MESA Press.

Linacre, J.M. (1994a). Constructing measurement with a many-facet Rasch model. In M. Wilson (Eds.), *Objective measurement: Theory into practice* (Vol. 2, pp 129-144). Norwood, NJ: Ablex.

Linacre, J.M. (2014). *A user's guide to FACETS Rasch-model computer programs.* 13 Temmuz 2014 tarihinde http://www.winsteps.com/a/facets-manual.pdf adresinden alınmıştır.

Linacre, J.M., Engelhard, G.Jr., Tatum, D.S., & Myford, C.M. (1994). Measurement with judges: Many-faceted conjoint measurement. *International Journal of Educational Research*, 21(6), 569-577.

Lister, R., Simon, B., Thompson, E., Whalley, J.L., & Prasad, C. (2006). Not seeing the forest for the trees: Novice programmers and the SOLO taxonomy. *ACM SIGCSE Bulletin*, 41(3), 118-122.

Liu, X. (2010). *Using and developing measurement instruments in science education: A Rasch modeling approach*. Charlotte, NC: Information Age Publishing.

Lottridge, S.M., Schulz, E.M., & Mitzel, H.C. (2012). Using automated scoring to monitor reader performance and detect reader drift in essay scoring. http://www.pacificmetrics.com/white-papers/NCME_2012_MonitoringReadersUsingAutomatedScoring.pdf adresinden alınmıştır.

Lucas, U., & Mladenovic, R. (2008). The identification of variation in students' understandings of disciplinary concepts: The application of the SOLO taxonomy within introductory accounting. *Higher Education*, 58(2), 257-283.

Lumley, T., & McNamara, T. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, 12(1), 54-71.

MacMillan, P.D. (2000). Classical, generalizability and multifaceted Rasch detection of interrater variability in large sparse data sets. *The Journal of Experimental Education*, 68(2), 167-190.

Maddrell, J.A. (2011). *Community of inquiry framework and learnıng outcomes*. Unpublished Doctoral Dissertation. Old Dominion University, Norfolk, Virginia, ABD.

Marcoulides, G.A. (2000). Generalizability theory. In H.E.A. Tinsley and S.D. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 527-551). San Diego: Academic Press

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

Matsuno, S. (2009). Self-peer-and teacher-assessments in Japanese university EFL writing classrooms. *Language Testing*, 26(1),75-100.

May, G.L. (2005, Ekim). *The effect of rater training on reducing social style bias in peer evaluation: A pilot study*. Paper Presented The Association for Business Communication Annual Convention, Irvine, California.

McBee, M.M., & Barnes, L.L.B. (1998) The generalizability of a performance assessment measuring achievement in eight-grade mathematics. *Applied Measurement in Education*, 11(2), 179-194.

McIntyre, P.N. (1993). *The importance and effectiveness of moderation training on the reliability of teachers' assessments of ESL writing samples*. Unpublished Master's Thesis, University of Melbourne, Melbourne.

McIntyre, R.M., Smith, D.E., & Hassett, C.E. (1984). Accuracy of performance ratings as affected by rater training and purpose of rating. *Journal of Applied Psychology*, 69(1), 147-156.

McNamara, T.F. (1996). *Measuring second language performance*. London and New York: Longman.

Mertler, C. (2001). Designing scoring rubrics for your classroom. *Practical Assessment, Research & Evaluation*, 7(25). 15 Ağustos 2013 tarihinde http://pareonline.net/getvn.asp?v=7&n=25 adresinden alınmıştır.

Messick, (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9),741-749.

Milli Eğitim Bakanlığı. (2007). *Matematik öğretmen klavuz kitabı*. Ankara: Devlet Kitapları Müdürlüğü.

Milli Eğitim Bakanlığı. (2009). *İlköğretim matematik dersi 6-8. sınıflar öğretim programı*. Ankara: Devlet Kitapları Müdürlüğü.

Milli Eğitim Bakanlığı. (2011). *Ortaöğretim matematik dersi (9, 10, 11 ve 12. sınıflar-haftalık 4 Saat) öğretim programı ve ortaöğretim matematik (10, 11 ve 12. sınıflar-haftalık 2 saat) dersi öğretim programı*. Ankara: Devlet Kitapları Müdürlüğü.

Milli Eğitim Bakanlığı. (2013). *Temel eğitimden ortaöğretime geçişle ilgili sıkça sorulan sorular*. 6 Kasım 2013 tarihinde http://www.meb.gov.tr/duyurular/duyurular2013/bigb/tegitimdenoogretimegecis/MEB_SSS_20_09_2013.pdf adresinden alınmıştır.

Milli Eğitim Bakanlığı. (2014a). *İlköğretim fen ve teknoloji 7. sınıf ders kitabı*. Ankara: Devlet Kitapları Müdürlüğü Basım Evi.

Milli Eğitim Bakanlığı. (2014b). *İlköğretim* sosyal bilgiler 6. Sınıf ders kitabı. Ankara: Devlet Kitapları Müdürlüğü Basım Evi.

Minogue, J., & Jones, G. (2009). Measuring the impact of haptic feedback using the SOLO taxonomy. *International Journal of Science Education*, 31(10), 1359-1378.

Mohd Nor, N., & Idris, N. (2010). Assessing students' informal inferential reasoning using SOLO taxonomy based framework. *Procedia Social and Behavioral Sciences*, 2(2), 4805-4809.

Montgomery, K. (2001). *Authentic assessment: A guide for elementary teachers*. New York: Longman.

Mooney E.S. (2002). A framework for characterizing middle school students' statistical thinking, *Mathematical Thinking and Learning*, 4(1), 23-63.

Moore, B.B. (2009). *Consideration of rater effects and rater design via signal detection theory*. Unpublished Doctoral Dissertation. Columbia University, New York.

Moskal, B.M. (2000). Scoring rubrics: What, when, how? *Practical Assessment, Research and Evaluation*, 7(3). 15 Ağustos 2013 tarihinde http://pareonline.net/getvn.asp?v=7&n=3 adresinden alınmıştır.

Moskal, B.M. & Leydens, J.A. (2000). Scoring rubric development: validity and reliability. *Practical Assessment, Research & Evaluation*, 7(10). 14 Ağustos 2013 tarihinde http://pareonline.net/getvn.asp?v=7&n=10 adresinden alınmıştır.

Mulqueen C., Baker D., & Dismukes, R.K. (2000, April) *Using multifacet Rasch analysis to examine the effectiveness of rater training*. Presented at the 15th Annual Conference for the Society for Industrial and Organizational Psychology (SIOP). New Orleans. 19 Eylül 2013 tarihinde http://www.air.org/files/multifacet_Rasch.pdf adresinden alınmıştır.

Myford, C.M., & Wolfe, E.W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, 4(4), 386-422.

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

Myford, C.M., & Wolfe, E.W. (2009). Monitoring rater performance over time: A framework for detecting differential accuracy and differential scale category use. *Journal of Educational Measurement*, 46(4), 371-389.

National Assessment Governing Board. U.S. Department of Education [NAGB]. (2002). *Mathematics framework for the 2003 national assessment of educational progress.* 26 Ekim 2013 tarihinde http://academic.wsc.edu/faculty/jebauer1/mat645/framework_03.pdf adresinden alınmıştır.

Nelson, N.W., & Van Meter, A.M. (2007). Measuring written language ability in narrative samples. *Reading & Writing Quarterly*, 23(3), 287-309.

Newman, D.A., Kinney, T.K., & Farr, J.L. (2005). Job performance ratings. In M. Hersen (Eds.), *Comprehensive handbook of psychological assessment, industrial and organizational assessment* (pp. 371-386). New Jersey: John Wiley & Sons, Inc.

Nitko, A.J. (2004). *Educational assessment of students*. Upper Saddle River, NJ: Pearson.

Noonan, L.E., & Sulsky, L.M. (2001) Impact of frame-of- reference and behavioral observation training on alternative training effectiveness criteria in a Canadian military sample. *Human Performance*, 14(1), 3-26.

Novick, M.R. (1966). The axioms and principle results of classical test theory. *Journal of Mathematical Psychology*, 3(1), 1-18.

Oaklef, M. (2009). Using rubrics to assess information literacy: An examination of methodology and ınterrater reliability. *Journal of the American Society for Information Science and Technology,* 60(5), 969-983.

Offir, B., Lev, Y., & Bezalel, R. (2008). Surface and deep learning processes in distance education: synchronous versus asynchronous systems. *Computers & Education*, 51(3), 1172-1183.

O'Neill, T.R., & Lunz, M.E. (1996, April). *Examining the invariance of rater and project calibrations using a multi-facet Rasch model*. Paper Presented at the Annual Meeting of the American Educational Research Association, New York.

Osterlind, S.J. (1998). *Constructing test items: Multiple-choice, constructed-response, performance, and other formats*. Norwell, MA: Kluwer Academic Publisher.

Ostini, R., & Nering, M.L. (2006). *Polytomous item response theory models*. Thousand Oaks, CA: SAGE Publications.

Öğrenci Seçme ve Yerleştirme Merkezi. (2013). *Açık uçlu sorularla deneme sınavının uygulanması*. 13 Ocak 2015 tarihinde http://www.osym.gov.tr/belge/1-19410/acik-uclu-sorularla-deneme-sinavinin-uygulanmasi-311201-.html adresinden alınmıştır.

Ömür, S. ve Erkuş, A. (2013). Dereceli puanlama anahtarıyla, genel izlenimle ve ikili karşılaştırmalar yöntemiyle yapılan değerlendirmelerin karşılaştırılması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, 28(2), 308-320.

Özder, H. (2012). Öğretmen adaylarının seçiminde uygulanan mülakat sınavının geçerlik ve güvenirliği: KKTC örneği. *Eğitim ve Bilim*, 37(166), 155-169.

Özmantar, M.F., Bingölbali, E. ve Akkoç, H. (2008, Mayıs). *İlköğretim sınıf öğretmenlerinin açık uçlu matematik soruları değerlendirme süreçleri*. VII. Ulusal Sınıf Öğretmenliği Eğitimi Sempozyumu'nda sunulmuş sözlü bildiri, Çanakkale, Türkiye.

Özmen Hızarcıoğlu, B. (2013). *Problem çözme sürecinde derecel puanlama anahtarı (rubrik) kullanımında puanlayıcı uyumunun incelenmesi*. Yayınlanmamış Yüksek Lisans Tezi, Abant İzzet Baysal Üniversitesi, Bolu, Türkiye.

Padiotis, I., & Mikropoulos, T.A. (2010). Using SOLO to evaluate an educational virtual environment in a technology education setting. *Educational Technology & Society*, 13(3), 233-245.

Pallant, J. (2005*). SPSS survival manual: A step by step guide to data analysis using SPSS for windows*. Australia: Australian Copyright.

Palm, T. (2008). Performance assessment and authentic assessment: A conceptual analysis of the literature. *Practical Assessment, Research & Evaluation*, 13(4), 1-11.

Parke, C.S. (2001). An approach that examines sources of misfit to improve performance assessment items and rubrics. *Educational Assessment* 7(3), 201-25.

Parlak, B. ve Doğan, N. (2014). Dereceli puanlama anahtarı ve puanlama anahtarından elde edilen puanların uyum düzeyleri. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, 29(2), 189-197.

Pegg, J., & Tall, D. (2005). The fundamental cycle of concept construction underlying various theoretical frameworks. *International Reviews on Mathematical Education (Zentralblatt für Didaktik der Mathematik)*, 37(6), 468-475.

Pellegrino, J.W., Baxter, G.P., & Glaser, R. (1999). Addressing the two disciplines problem: Linking theories of cognition and learning with assessment and ınstructional practice. *Review of Research in Education*, 24(1), 307-353.

Penny, J., Johnson, R.L., & Gordon, B. (2000). Using rating augmentation to expand the scale of an analytic rubric. *The Journal of Experimental Education*, 68(3), 269-287.

Perlman, C.C. (2003). *Performance assessment: Designing appropriate performance tasks and scoring rubrics*. North Carolina, USA.

Peter, F., & Alberto, B. (2013). Higher order thinking in chemistry curriculum and its asessment. *Thinking Skills and Creativity*, 10, 250-264.

Petkov, D., & Petkova, O. (2006). Development of scoring rubrics for IS projects as an assessment tool. *Issues in Informing Science and Information Technology*, 3, 499-510.

Popham, W.J. (1997). What's wrong-and what's right-with rubrics. *Educational Leadership*, 55(2), 72-75.

Pomplun, M., Capps, L., & Sundbye, N. (1998). Criteria teachers use to score performance items. *Educational Assessment*, 5(2), 95-110.

Prosser, M., Trigwell, K., & Waterhouse, F. (2000). Students' experiences of studying physics concepts: The effects of disintegrated perceptions and approaches. *European Journal of Psychology of Education*, 15(1),61-74.

Pulakos, E. (1991). Rater training for performance appraisal. In J.W. Jones, B. Steffy, D. Bray, B.D. Steffy., and D.W. Bray (Eds.), *Applying psychology in business: The handbook for managers and human resources professionals* (pp. 326-332). New York: Lexington Books.

Purpura, J.E. (2004). *Assessing grammar*. Cambridge University Press.

Ramineni, C. (2008). *Rater contrast effects in performance assessments using the medical licensure examination*. Unpublished Doctoral Dissertation, University of Delaware, Newark, Delaware, ABD.

Qinghui, G. (2013, June). *Many-facets analysis of overseas students' evaluation of teaching*. Paper Presented at the International Conference on Education Technology and Management Science, Nanjing, Jiangsu, China.

Ramsden, P. (2002). *Learning to teach in higher education*. London: Routledge

Randall, J., & Engelhard, G. (2009). Differences between teachers' grading practices in elementary and middle schools. *The Journal of Educational Research*, 102(3), 175-185.

Rasch, G. (1980). *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago: MESA Press.

Reckase, M.D. (2009). *Multidimensional item response theory*. New York: Springer.

Reckase, M.D., Ackerman, T.A., & Carlson, J.E. (1988). Building a unidimensional test using multidimensional items. *Journal of Educational Measurement*, 25(3), 193-203.

Reddy, M.Y. (2010). Design and development of rubrics to improve assessment outcomes. A pilot study in a master's level business program in india. *Quality Assurance in Education*, 19(1), 84-104.

Reddy, M.Y, & Andrade, H. (2010) A review of rubric use in higher education. *Assessment & Evaluation in Higher Education*, 35(4), 435-448.

Reeve, B.B., & Fayers, P. (2005). Applying item response theory modeling for evaluating questionnaire item and scale properties. In P. Fayers and R.D. Hays (Eds.), *Assessing Quality of Life in Clinical Trials: Methods of Practice* (pp. 55-73). Oxford University Press.

Richardson, J.C., Ertmer, P.A., & Lehman, A. (2007, April). *Peer feedback in online discussions: Can we impact students' critical thinking skills?* Paper Presented at the American Educational Research Association (AERA) Annual Conference. Chicago.

Riddle, E.J., & Smith, M. (2008). Developing and using rubrics in quantitative business courses. *The Coastal Business Journal*, 7(1), 82-95.

Robb Singer, N., & LeMahieu, P. (2011). The effect of scoring order on the independence of holistic and analytic scores. *Journal of Writing Assessment*, 4(1). 15 Temmuz 2014 tarihinde http://journalofwritingassessment.org/article.php?article=51 adresinden alınmıştır.

Roblyer, M.D., & Wiencke, W.R. (2003) Design and use of a rubric to assess and encourage interactive qualities in distance courses. *American Journal of Distance Education*, 17(2), 77-98.

Roch, S.G., & O'Sullivan, B.J. (2003). Frame of reference rater training issues: Recall, time and behavior observation training. *International Journal of Training and Development*, 7(2), 93-107.

Romagnano, L. (2001). The myth of objectivity in mathematics assessment. *Mathematics Teacher*, 94(1), 31-37.

Romberg, T.E., & Wilson, L.D. (1992). Issues related to development of authentic assessment system for school mathematics. In T.A. Romberg (Eds.), *Reform in school mathematics and authentic assessment* (pp.1-18). Albany: State University of New York Press.

Saal, F.E., Downey, R.G., & Lahey, M.A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, 88(2), 413-428.

Saito, H. (2008). EFL classroom peer assessment: Training effects on rating and commenting. *Language Testing*, 25(4), 553-581.

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing*, 25(4), 465-493.

Schmitt, N. (2012). *The oxford handbook of personnel assessment and selection*. New York: Oxford University Press.

Schuster, C. (2002). A mixture model approach to indexing rater agreement. *British Journal of Mathematical and Statistical Psychology*, 55(2), 289-303.

Schuster, C. (2004). A note on the interpretation of weighted kappa and its relations to other rater agreement statistics for metric scales. *Educational and Psychological Measurement*, 64(2), 243-253.

Scouller, K. (1998). The influence of assessment method on students' learning approaches: Multiple choice question examination versus assignment essay. *Higher Education*, 35(4), 453-472.

Scullen, S.E., Mount, M.K., & Goff, M. (2000). Understanding the latent structure of job performance ratings. *Journal of Applied Psychology*. 85(6), 956-970.

Sebok, S.S. Luu, K., & Klinger, D.A. (2013). Psychometric properties of the multiple mini-interview used for medical admissions: Findings from generalizability and Rasch analyses. *Advances in Health Sciences Education, Theory and Practice*,19(1), 71-84.

Semerci, Ç. (2011a). Doktora yeterlikler çerçevesinde öğretim üyesi, akran ve öz değerlendirmelerin Rasch ölçme modeliyle analizi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, 2(2), 164-171.

Semerci, Ç. (2011b). Mikro öğretim uygulamalarının çok yüzeyli Rasch ölçme modeli ile analizi. *Eğitim ve Bilim*, 36(161), 14-25.

Serow, P. (2007). Utilising the Rasch model to gain insight into students' understandings of class inclusion concepts in geometry. In J. Watson and K. Beswick (Eds.), *Mathematics: Essential research, essential practice*-Vol. 2 (pp. 651-666) Proceedings of the 30th Annual Conference of the Mathematics Education Research Group of Australasia, MERGA Inc.

Selden, S., Sherrier, T., & Wooters, R. (2012). Experimental study comparing a traditional approach to performance appraisal training to a whole-brain training method at C.B. Fleet laboratories. *Human Resource Development Quarterly*, 23(1), 19-34.

Shrout, P.E., & Fleiss, J.L. (1979). Intraclass correlation: Uses in assessing rater reliability. *Psychological Bullutein*, 86(2), 420-428.

Simkin, M.G., & Kuechler, W.L. (2005). Multiple-choice tests and student understanding: What is the connection. *Decision Sciences Journal of Innovative Education*, 3(1), 73-97.

Sinclair, A.L. (2000). *Differentiating rater accuracy training programs*. Unpublieshed Master Thesis, Virginia Polytechnic Institute and State University, Blacksburg, Virjinya, ABD.

Slack, F., Beer, M., Armitt, G., & Green,S. (2003). Assessment and learning outcomes: The evaluation of deep learning in an on-line course. *Journal of Information Technology Education*, 2, 305-317.

Spool, M.D. (1978). Training programs for observers of behavior: A review. *Personnel Psychology*, 31(4), 853-888.

Stecher, B. (2010). *Performance assessment in an era of standards-based educational accountability*. Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.

Sudweeks, R.R., Reeve, S., & Bradshaw, W.S. (2005). A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing*, 9(3), 239-261.

Şahinkarakaş, Ş. (2010). Raters' conceptions of a good writing and effect of training on their conceptions. *World Applied Sciences Journal*, 11(6), 695-699.

Tan, K.H.K., Tan, C.T., & Chua, J.S.M. (2008). Innovation in education: The "teacher, less, learn more" initiative in Singapore schools. In J. E. Larkley, & V.B. Maynhard (Eds.), *Innovation in education* (pp. 153-171). New York: Nova Science Publishers.

Tan, S. (2013). Validation of an analytic rating scale for writing: A Rasch modeling approach. *Iranian Journal of Language Testing*, 3(1), 1-10.

Tavşsancıl, E. (2010). *Tutumların ölçülmesi ve SPSS ile veri analizi*. Ankara: Nobel Yayın Dağıtım.

Tekin, H. (2009). *Eğitimde ölçme ve değerlendirme*. Ankara: Yargı Yayınevi.

Thompson, E. (2007). Holistic assessment criteria-applying SOLO to programming projects. In Mann, S. and Simon (Eds.), Proceedings of the Ninth Australasian Computing Education Conference (ACE2007). Ballarat, Victoria, Australia, Australian Computer Society Inc. 07 Ekim 2013 tarihinde http://crpit.com/confpapers/CRPITV66Thompson.pdf adresinden alınmıştır.

Thorndike, E.L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology*, 4(1), 25-29.

Tombari, M., & Borich, G. (1999). *Authentic assessment in the classroom: Applications and practice*. Upper Saddle River, NJ: Merrill.

Tuckman, B.W. (1991). Evaluating the alternative to multiple-choice testing for teachers. *Contemporary education*, 62(4), 299-300.

Tuna, A. (2011). *Trigonometri öğretiminde 5E öğrenme döngüsü modelinin öğrencilerin matematiksel düşünme ve akademik başarılarına etkisi*. Yayınlanmamış Doktora Tezi, Gazi Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara.

Turgut, M.F. ve Baykul, Y. (2012). *Eğitimde ölçme ve değerlendirme*. Ankara: Pegem Akademi Yayınları.

Turner, J. (2003). *Examining on art portfolio assessment using a many facet Rasch measurement model*. Yayınlanmamış Doktora Tezi, Boston College, Boston.

Urbina, S. (2004). *Essentials of psychological testing*. New Jersey: John Wiley & Sons. Inc.

Ülker, S. (2014). *İlköğretim T.C. inkılap tarihi ve Atatürkçülük 8. sınıf ders kitabı*. Ankara: Semih Ofset Matbaacılık Ltd. Şti.

Ünver, E. (2014). *İlköğretim fen ve teknoloji 8. sınıf ders kitabı*. Ankara: Dikey Yayıncılık.

Valenza, J.K. (2000). Students and teachers alike can benefit from rubrics. Interview with Dr. Andi Stix. *The Philadelphia Inquirer*. 23 Ağustos 2013 tarihinde http://www.uwstout.edu/soe/profdev/resources/upload/Students_and_Teachers_Alike_Can_Benefit_from_Rubrics.pdf adresinden alınmıştır.

Van Rossum, E.J., & Schenk, S.M. (1984). The relationship between learning conception, study strategy and learning outcome. *British Journal of Educational Psychology*, 54(1), 73-83.

Vanbelle, S., & Albert, A. (2009). A note on the linearly weighted kappa coefficient for ordinal scales. *Statistical Methodology*, 6(2), 157-163.

Venn, J.J. (2000). *Assessing students with special needs*. Upper Saddle River, NJ: Merrill/Prentice Hall.

Viera, A.J., & Garrett, J.M. (2005). Understanding interobserver agreement: The kappa statistic. *Family Medicine*, 37(5), 360-363.

Viswesvaran, C., Schmidt, F.L., & Ones, D.S. (2005). Is there a general factor in ratings of job performance? A meta-analytic framework for disentangling substantive and error influences. *Journal of Applied Psychology*, 90(1), 108-131.

Wadhwa, S. (2008). *A handbook of teaching and learning*. New Delhi, India: Sarup & Sons.

Wang, P. (2009). The inter-rater reliability in scoring composition. *English Language Teaching*, 2(3), 39-43.

Wang, B. (2010). On rater agreement and rater training. *English Language Teaching*, 3(1), 108-113.

Warrens, M.J. (2011). Cohen's kappa is a weighted average. *Statistical Methodology*, 8(6),473-484.

Weigle, S.C. (1998). Using FACETS to model rater training effects. *Language Testing*, 15(2) 263-287.

Weigle, S.C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing*, 6(2), 145-178.

Weigle, S.C. (2002). *Assessing writing*. New York: Cambridge University Press.

Weir, C.J. (2005). *Language testing and validation: An evidence-based approach*. New York: Palgrave MacMillan.

Weyers, M. (2006). *Teaching the FE Curriculum: Encouraging active learning in the classroom*. London: Continuum.

Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing*, 10(3), 305-319.

Wilson, M., & Case, H. (1997). An examination of variation in rater severity over time: A study in rater drift. In M. Wilson & G. Engelhard (Eds.), *Objective Measurement: Theory Into Practice* (Vol.5, pp. 113-134). Stamford, CT: Ablex.

Wiseman, C.S. (2008). *Investigating selected facets in measuring second language writing ability using holistic and analytic scoring methods*. Unpublished Doctoral Dissertation. Columbia University. New York, ABD.

Wiseman, C.S. (2012a). Rater effects: Ego engagement in rater decision-making. *Assessing Writing*, 17(3), 150-173.

Wiseman, C.S. (2012b). Comparison of the performance of analytic vs. holistic scoring rubrics to assess L2 writing. *Iranian Journal of Language Testing*, 2(1), 59-92.

Woehr, D. J. (1994). Understanding frame-of-reference training: The impact of training on the recall of performance information. *Journal of Applied Psychology*, 79(4), 525-534.

Woehr, D.J., & Huffuct, A.I. (1994). Rater training for performance appraisal. A qantitative review. *Journal of Occupational and Organizational Psychology*, 67(3), 189-205.

Woehr, D.J., & Roch, S. (2012). Supervisory performance ratings. In N. Schmitt (Eds.), *The Oxford handbook of personnel assessment and selection* (pp. 517-531). New York: Oxford University Press.

Wolf, K., & Stevens, E. (2007). The role of rubrics in advancing and assessing student learning. *The Journal of Effective Teaching*, 7(1), 3-14.

Wolfe, E.W. (2014). Methods for monitoring rating quality: Current practices and suggested changes. 16 Haziran http://researchnetwork.pearson.com/wp-content/uploads/Wolfe_MethodsForMonitoring_May2014-2.pdf adresinden alınmıştır.

Wolfe, E.W., & Dobria, L. (2008). Applications of the multifaceted Rasch model. In J.W. Osborne (Eds.), *Best practices in quantitative methods* (pp. 71-85). Los Angeles: Sage.

Wolfe, E.W., & McVay, A. (2010). Rater effects as a function of rater training context. 22 Ağustos 2013 tarihinde http://www.pearsonassessments.com/NR/rdonlyres/6435A0AF-0C12-46F7-812E-908CBB7ADDFF/0/RaterEffects_101510.pdf adresinden alınmıştır.

Wong, C.S.P. (2007). Views on the adoption and implementation of the SOLO Taxonomy. In S. Frankland (Eds.), *Enhancing teaching and learning through assessment* (pp. 4-15). The Netherlands: Springer.

Woodward, J., Monroe, K., & Baxter, J. (2001). Enhancing student achievement on performance assessments in mathematics. *Learning Disability Quarterly,* 24(1), 33-46.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement: Transactions of the Rasch Measurement SIG*, 8(3), 370.

Wynd, C.A., Schmidt, B., & Schaefer, M.A. (2003). Two quantitative approaches for estimating content validity. *Western Journal of Nursing Research*, 25, 508-518.

Yazıcı, N. (2013). *Başarının ölçülmesinde SOLO taksonomiye dayalı hazırlanan rubrik kullanımının etkisinin karşılaştırmalı olarak incelenmesi*. Yayınlanmamış Yüksek Lisans Tezi, Kahramanmaraş Sütçü İmam Üniversitesi, Sosyal Bilimler Enstitüsü, Kahramanmaraş.

Yıldırım, A. ve Şimşek H. (2011). *Sosyal bilimlerde nitel araştırma yöntemleri*. Ankara: Seçkin Yayınevi.

Yue, X. (2011). *Detecting rater centrality effect using simulation methods and Rasch measurement analysis.* Unpublished Doctoral Thesis, Virginia State University, Petersburg.

Zewotir, T. (2012). On employees' performance appraisal: The impact and treatment of the raters' effect. *South African Journal of Economic and Management Sciences,* 15(1), 44-54.

Zhu, X. (2009). *Assessing fit of item response models for performance assessments using bayesian analysis*. Unpublished Doctoral Thesis, University of Pittsburgh, Pittsburgh, ABD.

EKLER

Ek-1: Açık Uçlu Sorulardan Oluşan Matematik Başarı Testi

Değerli Öğrenciler, aşağıda yer alan matematik sorularını çözmeniz ve her bir soru için çözüme nasıl ulaştığınızı açıklamanız istenmektedir. Bu sorularla elde edilen bilgiler bilimsel bir araştırmada kullanılacaktır. Soruları çözerken, gerçek bir sınavdaymış gibi çözmeniz gerçeği yansıtan bilgiler elde edilebilmesi için son derece önemlidir. İlginize teşekkür ederim.

Arş. Gör. Mustafa İLHAN

Cinsiyetiniz: Kız ☐ Erkek ☐

Birinci Dönem Matematik Dersi Karne Notunuz: 1☐ 2☐ 3☐ 4☐ 5☐

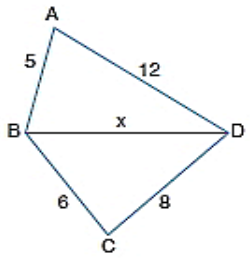
Birinci Dönem Yapılan Merkezi Ortak Sınavda Matematik Dersi Doğru Sayınız:…...

1. Esnaf Veli Bey, maliyeti x TL olan bir malı y TL'ye satmaktadır. x ile y arasında **"y = 8x-70"** bağıntısı vardır. Veli Bey'in bu malın satışından kar ettiği bilindiğine göre, malın maliyet fiyatı olan x için ne söyleyebilirsiniz.

Çözüm:

Problemi nasıl çözdüğünüzü açıklayınız:

2.



Yanda verilen ABD ve CBD üçgenleri için, |AB|=5 cm, |AD|=12 cm, |BC|=6 cm ve |CD|=8 cm olduğuna göre, |BD|= x uzunluğunun alabileceği değerlerin hangi aralıkta yer aldığını bulunuz.

Çözüm:

Problemi nasıl çözdüğünüzü açıklayınız:

LÜTFEN ARKA SAYFAYA GEÇİNİZ

3. p bir reel sayı olmak üzere; 2p ve p+6 ifadelerinden hangisi daha büyüktür?

Çözüm:

Problemi nasıl çözdüğünüzü açıklayınız:

4. A = {1, 2} şeklinde 2 elemanlı ve A U B = {1, 2, m, n} şeklinde 4 elemanlı bir küme olduğuna göre,

a) oluşturabileceğiniz farklı B kümelerini gösteriniz.
b) oluşturabileceğiniz farklı B kümelerinin sayısı ile A kümesinin eleman sayısı arasında nasıl bir ilişki olduğunu bulunuz.

Çözüm:

Problemi nasıl çözdüğünüzü açıklayınız:

5. İsmail'in çiftliğindeki 40 hayvanın 30'u inek, 10'u koyundur. Kenan'ın çiftliğindeki 60 hayvanın 40'ı inek, 20'si koyundur. İsmail'in ve Kenan'ın çiftliğinden birer hayvan seçiliyor. Hem İsmail'in hem de Kenan'ın çiftliğinden seçilen hayvanın koyun olma olasılığını bulunuz.

Çözüm:

Problemi nasıl çözdüğünüzü açıklayınız:

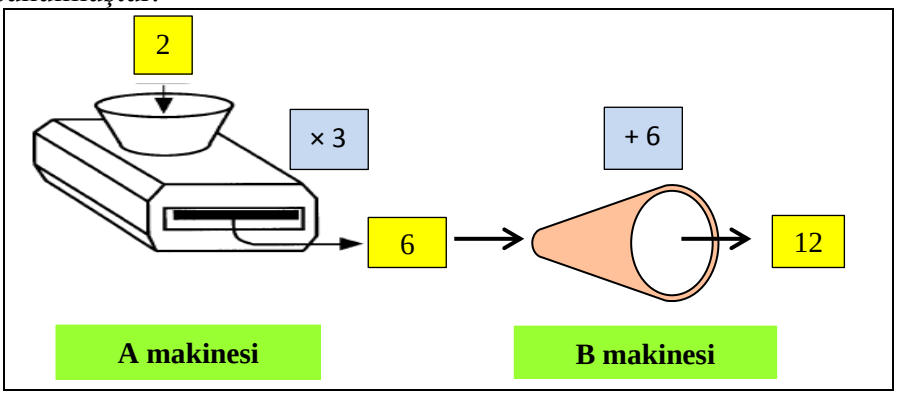
LÜTFEN DİĞER SAYFAYA GEÇİNİZ

6. a ve b birer reel sayı, a+b=8 ve b < a olduğuna göre, b'nin alabileceği değerler için ne söyleyebilirsiniz?

Çözüm:

Problemi nasıl çözdüğünüzü açıklayınız:

7. A ve B iki matematiksel denklem makinesi olmak üzere, A makinesine bir sayı konulduğunda makine bu sayıyı 3 ile çarparak değiştirmektedir. Daha sonra, A makinesinden çıkan sayı B makinesine girmekte ve B makinesi bu sayıya 6 ekleyerek değiştirmektedir. Aşağıda A makinesine 2 sayısının konulduğu bir örnek sunulmuştur.



Örnekte görüldüğü gibi A makinesi, konulan sayıyı 3 ile çarpmış ve A makinesinden çıkan sayı 6 olmuştur. Daha sonra bu sayı B makinesine girmiştir. B makinesi sayıyı 6 ekleyerek değiştirmiş ve sayı B makinesinden 12 olarak çıkmıştır. A ve B makinelerinin çalışma prensiplerine göre, aşağıdaki soruları cevaplayınız.

a) A makinesine 4 sayısı konulduğunda B makinesine girecek sayı kaç olur?
b) A makinesine sırasıyla 7 ve 12 sayıları konulduğunda B makinesinden çıkan sayılar kaç olur?
c) A makinesine x sayısı konulduğunda B makinesinden çıkan sayı y olmaktadır. Buna göre, y sayısını x cinsinden ifade ediniz.
d) A makinesine giren sayı ile B makinesinden çıkan sayı arasındaki eşitlik değişmeyecek şekilde A ve B makineleri için başka birer çalışma kuralı belirleyiniz.

Çözüm:

Problemi nasıl çözdüğünüzü açıklayınız:

LÜTFEN ARKA SAYFAYA GEÇİNİZ

8. Ali ve Ayşe oynadıkları oyunda, ellerindeki eşit boydaki kürdanlar ile yan yana evler yapmaya çalışmaktadır. Aşağıda Ali ve Ayşe'nin bu oyunu oynarken yaptıkları evler görülmektedir. Buna göre,

a) 9 ev için gerekli olan kürdan sayısını hesaplayınız.
b) 42 ev için gerekli olan kürdan sayısı 169 ise, 43 ev için gerekli olan kürdan sayısı kaçtır?
c) Ev sayısı ile kürdan sayısı arasındaki ilişkiyi cebirsel olarak ifade ediniz.
d) Ali ve Ayşe "*beşgen*" şeklindeki evler yerine yine yan yana olacak şekilde farklı bir geometrik şekilden oluşan evler yapmak istiyor. Ayşe ve Ali'ye yardımcı olmak için farklı bir geometrik şekil belirleyiniz. Belirlediğiniz geometrik şekilden yapılan evler için, kürdan sayısı ile ev sayısı arasındaki ilişkiyi cebirsel olarak ifade ediniz.

Çözüm:

Problemi nasıl çözdüğünüzü açıklayınız:

İLGİNİZE TEŞEKKÜR EDERİM.

Ek-2: Açık Uçlu Matematik Sorularının Puanlanmasında Kullanılan Standart Rubrikler

Soru1:

Puanlama Ölçütleri	
3 puan	Problem tam olarak anlaşılmış ve uygun strateji kullanılarak doğru cevaba ulaşılmıştır. -Öğrenci Veli Bey'in kar edebilmesi için satış fiyatı olan y'nin alış fiyatı olan x'ten büyük olması gerektiğinin farkındadır. y=8x-70 bağıntısını kullanarak 8x-70>x eşitsizliğini elde edebilmiştir. Bu eşitsizliği hatasız bir biçimde çözerek y>x olması için x>10 olmalı şeklinde doğru cevaba ulaşmıştır. Veli Bey'in kar zarar durumuna ilişkin yapılan işlemler açık, ayrıntılı ve örnek cevap niteliğindedir.
2 puan	Problem büyük ölçüde anlaşılmıştır. -Çözüm genel olarak doğru olup yalnızca küçük hatalar bulunmaktadır. Veli Bey'in kar edebilmesi için y>x olmalıdır şeklinde uygun strateji ile çözüme başlanmış ve y=8x-70 bağıntısı kullanarak strateji 8x-70>x şeklinde devam ettirmiştir. Ancak 8x-70>x eşitsizliğini çözerken küçük işlem hatalarından veya anlaşılamayan nedenlerden dolayı sonuca ulaşamamış ya da yanlış sonuca ulaşmıştır. -Öğrenci satış fiyatının alış fiyatından büyük olması için x>10 olmalı şeklinde doğru sonuca ulaşmıştır. Ancak problemi nasıl çözdüğüne yönelik yeterli açıklama yapmamıştır.
1 puan	Problem kısmen anlaşılmıştır. -Öğrenci yalnızca uygun strateji ile başlangıç yapmış ama devamını getirememiştir. Örneğin, "Veli Bey'in kar edebilmesi için y>x olmalıdır" şeklinde uygun strateji ile başlangıç yapılmış olabilir. Ancak strateji devam ettirilerek "y yerine 8x-70 yazılıp" 8x-70>x eşitsizliği kurulamamıştır. -Uygun strateji ile başlangıç yapılarak y>x eşitsizliği kurulmuştur. Ancak, y>x eşitliği kurulduktan sonra problemin çözümüne yönelik doğru işlemler yapılamamıştır. Yapılan işlemlerde önemli hatalar bulunmaktadır.
0 puan	Öğrenci problemi anlamamıştır. -"Problemde hem y hem de x gibi iki tane bilinmeyen var, ne yapacağımı bilmiyorum" gibi ifadeler kullanılmıştır. -Öğrenci problemin çözümünü bulmaya yönelik herhangi bir işlem yapmamıştır. -Öğrenci kar zarar durumunu çözmeye yönelik herhangi bir işlem yapmamıştır. Sadece "Esnaf Veli Bey malı x TL'ye alıp y TL'ye satmaktadır" gibi problemin tekrarı niteliğindeki ifadeler kullanılmıştır. -Öğrenci Veli Bey'in kar-zarar durumuna ilişkin doğru cevaba ulaştırması mümkün olmayan yanlış eşitlikler kurmuş, yanlış ifadeler ve stratejiler kullanılmıştır.

Soru 2:

Puanlama Ölçütleri	
3 puan	Problem tam olarak anlaşılmıştır. -Üçgen eşitsizliğinden yararlanılarak doğru cevaba ulaşılmıştır. Öğrenci x uzunluğunu hesaplayabilmek için her iki üçgende de üçgen eşitsizliği kurması ve daha sonra bu eşitsizliklerden elde edilen çözüm kümelerinin kesişimini alması gerektiğinin farkındadır. $7<x<17$ ve $2<x<14$ eşitsizliklerine ulaşılmış, bu eşitsizliklerin kesişimini alınarak $7<x<14$ şeklinde doğru cevap bulunmuştur. x uzunluğunun hangi aralıkta yer aldığına ilişkin yapılan işlemler açık, ayrıntılı ve örnek cevap niteliğindedir.
2 puan	Problem büyük ölçüde anlaşılmıştır. -Çözüm genel olarak doğru olup yalnızca küçük hatalar içermektedir. Örneğin, öğrenci her iki üçgende de üçgen eşitsizliğini kullanarak, ABD ve CBD üçgenlerinde x'in alabileceği değerlerin hangi aralıkta yer aldığını elde etmesi ve daha sonra bu değerlerin kesişimini alması gerektiğinin farkındadır. Ancak, küçük işlem hatalarından veya anlaşılmayan nedenlerden dolayı sonuca ulaşılamamış veya yanlış sonuca ulaşılmıştır. -Öğrenci $7<x<14$ şeklinde doğru cevaba ulaşmıştır. Ancak problemi nasıl çözdüğüne ilişkin yeterli açıklama yoktur.
1 puan	Problem kısmen anlaşılmıştır. -Problemin çözümüne yönelik uygun strateji olan üçgen eşitsizliği ile probleme başlangıç yapılmıştır. -Öğrenci, "Bir üçgenin herhangi kenarı; diğer iki kenarın farkından küçük, toplamından büyük olmaz" şeklindeki üçgen eşitsizliği kuralını verilen üçgenlere uygulaması gerektiğinin farkındadır. Ancak devamını getirememiştir. -Uygun strateji ile başlangıç yapmanın dışında problemin çözümüne yönelik doğru işlemler yapılamamıştır. Yapılan işlemlerde önemli hatalar bulunmaktadır.
0 puan	Problem anlaşılmamıştır. -Öğrenci "Problemde iki tane üçgen var, x'in alacağı değerleri hangi üçgene göre hesaplamalıyım" gibi ifadeler kullanmıştır. -Öğrenci x uzunluğunun hangi aralıkta yer aldığını bulmaya yönelik herhangi bir işlem yapmamıştır. -Öğrenci "Şekilde iki tane üçgen var, her iki üçgenin de iki kenarının uzunluğu biliniyor. Ancak üçüncü kenarının uzunluğu bilinmiyor" gibi problemin tekrarı niteliğindeki ifadeler kullanılmıştır. -Öğrenci, x uzunluğunun hangi aralıkta yer aldığına dair doğru cevaba ulaştırması mümkün olmayan yanlış strateji ile başlangıç yapmıştır

Soru 3:

Puanlama Ölçütleri	
3 puan	Problem tam olarak anlaşılmıştır. -p'nin alacağı değerlere göre iki ifadeden hangisinin daha büyük olduğunu bulabilmek için 2p ve p+6 ifadeleri arasında eşitlik ve eşitsizlik bağıntıları kurularak doğru cevaba ulaşılmıştır. Öğrencinin verdiği cevap; aşağıdaki gibi açık, anlaşılır ve örnek yanıt niteliğindedir. 2p>p+6 eşitsizliği p>6 için doğrudur. 2p=p+6 ifadesi p=6 için doğrudur. p+6>2p ifadesi p<6 için doğrudur.
2 puan	Problem büyük ölçüde anlaşılmıştır. -Çözüm genel olarak doğru olup yalnızca küçük hatalar bulunmaktadır. Öğrenci 2p>p+6, 2p=p+6 ve p+6>2p gibi bağıntılardan yararlanmıştır. Ancak küçük işlem hatalarından ya da anlaşılmayan nedenlerden dolayı eşitlik ve eşitsizliklerin çözüm kümesini bulmaya yönelik işlemleri sonuçlandıramamış veya yanlış sonuçlandırmıştır. -Öğrenci p=6 için iki ifade eşittir. p>6 için 2p ve p<6 için p+6 ifadesi daha büyüktür şeklinde doğru cevaba ulaşmıştır. Ancak problemi nasıl çözdüğüne ilişkin yeterli açıklama yoktur.
1 puan	Problem kısmen anlaşılmıştır. -Problemin çözümüne yönelik olarak 2p ve p+6 ifadeleri arasında eşitlik ya da eşitsizlik kurmak gibi uygun stratejiler ile probleme başlangıç yapılmıştır. Ancak devamını getirilememiştir. -Problemi çözmek için uygun strateji ile başlangıç yapmanın dışında problemin çözümüne yönelik doğru işlemler yapılamamıştır. Yapılan işlemlerde önemli hatalar bulunmaktadır.
0 puan	Problem anlaşılmamıştır. -Öğrenci "2p ve p+6 ifadelerinin her ikisi de bilinmiyor, hangisinin büyük olduğunu bulmak mümkün değil" gibi ifadeler kullanmıştır. -Öğrenci, 2p ve p+6 ifadelerinden hangisinin daha büyük olduğunu bulmaya yönelik herhangi bir işlem yapmamıştır. -Öğrenci, 2p ve p+6 ifadelerinden hangisinin daha büyük olduğunu bulmamız isteniyor gibi problemin tekrarı niteliğindeki ifadeler kullanılmıştır. -Öğrencinin, 2p ve p+6 ifadelerinden hangisinin daha büyük olduğunu bulmaya yönelik olarak kullandığı strateji yanlıştır.

Soru 4:

Puanlama Ölçütleri	
3 puan	Problem tam olarak anlaşılmıştır. -Öğrenci AUB kümesinin A kümesinden farklı elemanlarının mutlaka B kümesinin elemanları arasında olması gerektiğinin farkındadır. A kümesinin her bir alt kümesine "m, n" elemanlarını ekleyerek farklı B kümelerini elde edebilir ve bu kümelerin sayısının A kümesinin alt küme sayısına eşit olduğu sonucuna ulaşabilir. Öğrencinin, farklı B kümelerini göstermek ve bu kümelerin sayısı ile A kümesinin eleman sayısı arasındaki ilişki bulmak için yaptığı işlemler açık, ayrıntılı ve örnek yanıt niteliğindedir.
2 puan	-Problem büyük ölçüde anlaşılmıştır. -Çözüm genel olarak doğru olup yalnızca küçük hatalar içermektedir. Öğrenci, AUB kümesinin A kümesinden farklı elemanlarının B kümesinin elemanları arasında yer aldığını bilir. Ayrıca öğrenci, m ve n elemanlarının yanında A kümesinin elemanlarının da B kümesinin elemanı olabileceğinin farkındadır. Ancak işlemlerini sonuçlandıramaz veya yanlış sonuçlandırır. -Öğrenci; "{m, n}, {m, n, 1}, {m, n, 2} ve {m, n, 1, 2} olmak üzere A kümesinin alt küme sayısına eşit olan 4 farklı B kümesi oluşturulabilir" şeklinde doğru cevap verir. Ancak doğru cevaba nasıl ulaştığına ilişkin yeterli açıklamayı yapamaz.
1 puan	Problem kısmen anlaşılmıştır. -Öğrenci; ilk olarak AUB kümesinin A kümesinden farklı elemanlarını belirlemek gibi uygun bir strateji ile başlangıç yapar, ancak devamını getiremez. -Öğrenci uygun strateji ile başlangıç yapmanın dışında oluşturulabilecek farklı B kümelerini bulmaya yönelik doğru işlemler yapamamıştır. Yapılan işlemlerde önemli hatalar bulunmaktadır.
0 puan	Problem anlaşılmamıştır. -Öğrenci problemi tamamen boş bırakmış, çözüme yönelik herhangi bir işlem yapmamıştır. -Öğrenci B={3, 4} gibi yalnızca yanlış sonuçlar yazmıştır. -Öğrenci problemin çözüme yönelik tamamen yanlış stratejiler kullanmıştır. Kullanılan stratejinin doğru sonuca ulaştırması mümkün değildir. -Öğrenci, "A kümesi ile AUB kümesi verilmiş" gibi problemde yer alan bilgilerin tekrarı niteliğindeki ifadeler kullanmıştır.

Soru 5:

Puanlama Ölçütleri	
3 puan	Problem tam olarak anlaşılmıştır. -Öğrenci ilk olarak, İsmail'in ve Kenan'ın çiftliğinden seçilen birer hayvanın koyun olması olasılığını hesaplamış ve daha sonra bulduğu değerleri birbiri ile çarparak doğru sonuca ulaşmıştır. Öğrencinin, hem İsmail'in hem de Kenan'ın çiftliğinden seçilen hayvanın koyun olma olasılığını bulmak için yaptığı işlemler açık, ayrıntılı ve örnek yanıt niteliğindedir.
2 puan	Problem büyük ölçüde anlaşılmıştır. -Çözüm genel olarak doğru olup yalnızca küçük hatalar içermektedir. Öğrenci; İsmail'in ve Kenan'ın çiftliğinden seçilen birer hayvanın koyun olması olasılığını hesaplaması ve daha sonra elde etiği değerleri çarpması gerektiğinin farkındadır. Ancak, öğrenci yaptığı küçük işlem hatalarından veya anlaşılmayan nedenlerden dolayı işlemleri sonuçlandıramamış ya da yanlış sonuçlandırmıştır. -Öğrenci istenilen olasılık değerini doğru bir şekilde hesaplamıştır. Ancak problemi nasıl çözdüğüne ilişkin yeterli açıklama yoktur.
1 puan	Problem kısmen anlaşılmıştır. -Öğrenci ilk olarak, her bir çiftlikte koyun sayısının toplam hayvan sayısına oranını almak gibi doğru strateji ile çözüme başlangıç yapar ancak devamını getiremez. -Uygun strateji ile başlangıç yapmanın dışında problemin çözümüne yönelik doğru işlemler yapılmamıştır. Yapılan işlemlerde önemli hatalar bulunmaktadır.
0 puan	Problem anlaşılmamıştır. -Öğrenci istenilen olasılık hesabını bulmaya yönelik hiçbir işlem yapmamıştır. -Öğrenci problemin çözüme yönelik tamamen yanlış stratejiler kullanmıştır. Kullanılan stratejinin problemin çözümüne yönelik doğru sonucuna ulaştırması mümkün değildir. -Öğrenci "Çiftliklerden seçilen hayvanın koyun olma olasılığı olduğu gibi inek olma olasılığı da vardır" şeklinde problemin çözümüne yönelik herhangi bir yarar sağlamayan ya da "İsmail ve Kenan'ın çiftliğinden birer hayvan seçilmesi durumunda, her iki çiftlikten de seçilen hayvanın koyun olması olasılığını bulmamız isteniyor" gibi problemin tekrarı niteliğindeki ifadeler kullanmıştır.

Soru 6:

Puanlama Ölçütleri	
3 puan	Problem tam olarak anlaşılmıştır. -Öğrenci a+b=8 eşitliğinden yararlanarak b<a eşitsizliğinde a yerine 8-b yazmak gibi uygun bir strateji ile başlangıç yapar. b<8-b eşitsizliğini çözüp b<4 şeklinde doğru cevaba ulaşır. Öğrencinin, b'nin alabileceği değerleri bulmak için yaptığı işlemler açık, ayrıntılı ve örnek yanıt niteliğindedir.
2 puan	Problem büyük ölçüde anlaşılmıştır. -Çözüm genel olarak doğru olup yalnızca küçük hatalar bulunmaktadır. Öğrenci a+b=8 eşitliğinden yararlanarak b<a eşitsizliğinde a yerine 8-b yazmak gibi uygun bir strateji ile başlangıç yapar ve b<8-b eşitsizliğini çözmeye çalışarak bu stratejiyi devam ettirir. Ancak, işlemleri sonuçlandıramaz veya yanlış sonuçlandırır. -Öğrenci b<4 şeklinde doğru cevaba ulaşmıştır. Ancak problemi nasıl çözdüğüne ilişkin yeterli açıklama yoktur.
1 puan	Problem kısmen anlaşılmıştır. -Öğrenci; a+b=8 eşitliğinden yararlanarak; b<a eşitsizliğini yalnızca a veya yalnızca b cinsinden yazmak ya da a+b=8 eşitliğinde ilk aşamada a ve b'yi eşit kabul etmek gibi uygun strateji ile başlangıç yapar. Ancak devamının getiremez. -Öğrenci uygun strateji ile başlangıç yapmanın dışında problemin çözümüne yönelik doğru işlemler yapamamıştır. Yapılan işlemlerde önemli hatalar bulunmaktadır.
0 puan	Problem anlaşılmamıştır. -Öğrenci b'nin alabileceği değerleri bulmaya yönelik hiçbir işlem yapmamıştır. -Öğrenci "Hem a hem de b bilinmiyor, bu nedenle b'nin alabileceği değerler hakkında bir şey söyleyemeyiz" gibi ifadeler kullanmıştır. -Öğrenci problemin çözüme yönelik tamamen yanlış stratejiler kullanır. Kullanılan stratejinin problemin çözümüne yönelik doğru sonucuna ulaştırması mümkün değildir. -Öğrenci, "a+b'yi ve b'nin a'dan küçük olduğunu biliyoruz" şeklinde problemde verilenlerin tekrarı niteliğindeki ifadeler kullanmıştır.

Soru 7:

Puanlama Ölçütleri	
4 puan	Problem tam olarak anlaşılmıştır. Öğrenci tüm alt amaçlarda uygun strateji kullanarak doğru cevaba ulaşmıştır. Yaptığı çözümü iyi bir biçimde organize etmiş ve cevaba nasıl ulaştığını açıklamıştır. Öğrencinin yanıtı doğru, açık, ayrıntılı ve örnek cevap niteliğindedir.
3 puan	Problem büyük ölçüde anlaşılmıştır. -Öğrenci, problemin alt amaçlarına uygun stratejiler kullanmış ve bu stratejileri uygun işlemlerle devam ettirmiştir. Problemi nasıl çözdüğünü açıklamıştır. Ancak yaptığı küçük işlem hataları veya anlaşılmayan nedenlerden dolayı sonuca ulaşamamış ya da yanlış sonuca ulaşmıştır. -Öğrenci, problemin tüm alt amaçlarında uygun stratejiler ile başlamış ve doğru cevaba ulaşmıştır. Ancak sonuca nasıl ulaştığına dair yeterli açıklama yoktur.
2 puan	Problem kısmen anlaşılmıştır. -Öğrenci problemin alt amaçlarından sadece birinde uygun strateji kullanıp doğru sonuca ulaşmıştır ve sonuca nasıl ulaştığı anlaşılmaktadır. -Öğrenci problemin birden fazla alt amacında uygun stratejiyi seçmiş ancak bu stratejiler alt amaçların hiçbirinde devam ettirilememiştir. -Öğrenci birden fazla alt amaçta sadece doğru sonucu yazmıştır. Ancak öğrencinin sonuca nasıl ulaştığı açık değildir.
1 puan	Problem çok az anlaşılmıştır. -Öğrenci problemin alt amaçlarından yalnızca biri için uygun strateji belirlemiştir. Ancak bu stratejiyi sonuçlandıramamış veya yanlış sonuçlandırmıştır. -Öğrenci alt amaçlardan yalnızca birine doğru cevap vermiştir. Ancak doğru cevabı yazdığı bu alt amaçta cevaba nasıl ulaştığı açık değildir.
0 puan	-Öğrencinin problemi anladığına dair izler bulunmamaktadır. Öğrenci problemi hiç anlamamış veya yanlış anlamıştır. -Çözümü bulmaya yönelik hiçbir çalışma yapılmamıştır. -Öğrenci, yalnızca yanlış sonucu yazmıştır. -Öğrenci, "A ve B iki matematiksel denklem makinesidir" gibi sadece problemdeki verileri kopyalamıştır. -Öğrenci, problemin alt amaçlarının hiçbirinde uygun strateji kullanmamıştır. -Öğrenci, problemin alt amaçlarının tümünde çözüm ile ilgili olmayan yanlış ifadeler kullanmıştır.

Soru 8:

Puanlama Ölçütleri	
4 puan	Problem tam olarak anlaşılmıştır. Öğrenci tüm alt amaçlarda uygun strateji kullanarak doğru cevaba ulaşmıştır. Yaptığı çözümü iyi bir biçimde organize etmiş ve cevaba nasıl ulaştığını açıklamıştır. Öğrencinin yanıtı doğru, açık, ayrıntılı ve örnek cevap niteliğindedir.
3 puan	Problem büyük ölçüde anlaşılmıştır. -Öğrenci, problemin alt amaçlarına uygun stratejiler kulanmış ve bu stratejileri uygun işlemlerle devam ettirmiştir. Problemi nasıl çözdüğünü açıklamıştır. Ancak yaptığı küçük işlem hataları veya anlaşılmayan nedenlerden dolayı sonuca ulaşamamış ya da yanlış sonuca ulaşmıştır. -Öğrenci, problemin tüm alt amaçlarında uygun stratejiler ile başlamış ve doğru cevaba ulaşmıştır. Ancak sonuca nasıl ulaştığına dair yeterli açıklama yoktur.
2 puan	Problem kısmen anlaşılmıştır. -Öğrenci problemin alt amaçlarından sadece birinde uygun strateji kullanıp doğru sonuca ulaşmıştır ve sonuca nasıl ulaştığı anlaşılmaktadır. -Öğrenci problemin birden fazla alt amacında uygun stratejiyi seçmiş ancak bu stratejiler alt amaçların hiçbirinde devam ettirilememiştir. -Öğrenci birden fazla alt amaçta sadece doğru sonucu yazmıştır. Ancak öğrencinin sonuca nasıl ulaştığı açık değildir.
1 puan	Problem çok az anlaşılmıştır. -Öğrenci problemin alt amaçlarından yalnızca biri için uygun strateji belirlemiştir. Ancak bu stratejiyi sonuçlandıramamış veya yanlış sonuçlandırmıştır. -Öğrenci alt amaçlardan yalnızca birine doğru cevap vermiştir. Ancak doğru cevabı yazdığı bu alt amaçta cevaba nasıl ulaştığı açık değildir.
0 puan	- Öğrencinin problemi anladığına dair izler bulunmamaktadır. Öğrenci problemi hiç anlamamış veya yanlış anlamıştır. - Çözümü bulmaya yönelik hiçbir çalışma yapılmamıştır. - Öğrenci, yalnızca yanlış sonucu yazmıştır. - Öğrenci, “Ali ve Ayşe kürdanlardan beşgen şeklindeki evler yapıyorlar” gibi sadece problemdeki verileri kopyalamıştır. -Öğrenci, problemin alt amaçlarının hiç birinde uygun strateji kullanmamıştır. -Öğrenci, problemin alt amaçlarının tümünde çözüm ile ilgili olmayan yanlış ifadeler kullanmıştır.

Ek-3: Açık Uçlu Matematik Sorularının Puanlanmasında Kullanılan SOLO Taksonomisine Dayalı Rubrikler

Soru 1:

Puanlama Ölçütleri	
3 puan İlişkisel Yapı	Öğrenci x=10 TL'yi kritik değer olarak belirler. x'in 10 TL'den küçük olduğu durumlarda Veli Bey'in zarar ettiği, x=10 TL olması durumunda Veli Bey'in ne kar ne zarar ettiği ve x'in 10 TL'den büyük olması halinde Veli Bey'in kar ettiği şeklinde tutarlı bir sonuca ulaşmıştır.
2 puan Çok Yönlü Yapı	Öğrenci x'in değişken olduğunun farkındadır. x'e birden fazla sayıda değer vererek Veli Bey'in kar zarar durumunu belirlemeye çalışır. Ancak, Veli Bey'in malın satışından kar edebilmesi için x'in en az kaç TL olması gerektiğini bulamaz. x=10 TL'nin altındaki ve üstündeki değerler için Veli Bey'in kar ya da zarar durumunun değişeceğinin farkında değildir.
1 puan Tek Yönlü Yapı	Öğrenci değişken kavramının farkındadır. Ancak, problemi tek bir yönü ile ele aldığından x'e yalnızca bir değer vererek soruyu çözmeye çalışır. Örneğin, öğrenci "x=9 için y=72-70=2 olur ve y<x olduğundan Veli Bey zarar eder" gibi x'in tek bir değeri için yorum yapar.
0 puan Yapı Öncesi	Öğrenci, "Bilmiyorum", "Esnaflık bazen kar bazen de zarar ettiren bir meslektir" veya "y=8x-70 ifadesinde 8 tane x olduğu için Veli Bey kar eder" gibi ifadeler kullanabilir. Öğrenci değişken kavramının farkında değildir. Öğrenci, "x'in alabileceği değerlerden bağımsız olarak y daima x'ten daha büyüktür" gibi kavram yanılgılarına sahiptir.

Soru 2:

Puanlama Ölçütleri	
3 puan İlişkisel Yapı	Öğrenci hem ABD hem de CBD üçgeninde üçgen eşitsizliğini uygulayarak, 7<x<17 ve 2<x<14 eşitsizliklerinin her ikisine de elde edebilir. Daha sonra elde ettiği bu eşitsizlikleri tutarlı bir biçimde bir araya getirerek 7<x<14 sonucuna ulaşabilir.
2 puan Çok Yönlü Yapı	Öğrenci hem ABD hem de CBD üçgeninde üçgen eşitsizliğini uygulayarak, 7<x<17 ve 2<x<14 eşitsizliklerinin her ikisine de ulaşabilir. Ancak elde ettiği bu eşitsizlikleri birleştirerek anlamlı bir bütün oluşturamaz.
1 puan Tek Yönlü Yapı	Öğrenci, ABD ve CBD üçgenlerinden yalnızca biri için üçgen eşitsizliğini uygulayabilir. Dolayısıyla, öğrenci 7<x<17 ve 2<x<14 aralıklarından yalnızca birine ulaşabilir.
0 puan Yapı Öncesi	Öğrenci, "Şekilde 2 tane üçgen var" veya "x uzunluğunu bulmak çok zor" gibi çözümle ilgisi olmayan ifadeler kullanır. Yaptığı çözüm tamamen yanlıştır.

Soru 3:

Puanlama Ölçütleri	
3 puan İlişkisel Yapı	Öğrenci p=6'yı kritik değer olarak belirleyip 6'dan küçük ve büyük değerler için farklı durumlar oluşacağını kestirebilir. Öğrenci; p=6 için 2p ve p+6 ifadelerinin eşit olduğu, 6'dan büyük değerler için 2p ifadesinin, 6'dan küçük değerler için ise p+6 ifadesinin daha büyük olduğu şeklinde tutarlı bir sonuca ulaşmıştır.
2 puan Çok Yönlü Yapı	Öğrenci p'nin bir değişken olduğunun farkındadır. p değişkenine birden fazla değer vererek problemi çözmeye çalışır. Burada öğrenci problemi cevaplarken p'ye farklı değerler vererek yorum yapabilse de; olası bütün durumları göz önünde bulunduramaz. Özellikle p=6 için iki ifadenin birbirine eşit olduğunun, 6'dan küçük ve büyük değerler için farklı durumlar oluşacağının farkında değildir. Örneğin, öğrenci "p=2 için p+6 büyükken; p=10 için 2p büyük olur. Dolayısıyla bazen 2p, bazen de p+6 daha büyüktür" gibi ifadeler kullanır.
1 puan Tek Yönlü Yapı	Öğrenci p'ye tek bir değer vererek soruyu çözmeye çalışmıştır. Burada öğrenci, değişken kavramının farkındadır. Ancak, problemi tek bir yönü ile ele alır ve p'ye sadece bir değer vererek soruyu çözmeye çalışır. Örneğin öğrenci, p=3 için 2p=6 ve p+6=9 olur. Dolayısıyla p+6 ifadesi 2p'den büyüktür şeklinde bir cevap verir.
0 puan Yapı Öncesi	Öğrenci problemi anlamakta zorlanır. Sorunun çözümü ilgili olmayan cevaplar verir. Öğrencinin, değişken kavramı hakkında bir fikri olmadığından, öğrenci p+6=7 gibi benzer olmayan terimleri toplayabilir ya da 2p ve p+6 ifadelerinde p'ye farklı değerler verebilir.

Soru 4

Puanlama Ölçütleri	
3 puan İlişkisel Yapı	Öğrenci oluşturulabilecek farklı B kümelerini göstermekle kalmaz. Oluşturulabilecek farklı B kümelerinin sayısı ile A kümesinin eleman sayısı arasındaki ilişkiyi de açıklayabilir. Yani öğrenci hem oluşturabilecek farklı B kümelerini gösterebilir hem de bu kümelerin sayısının A kümesinin alt küme sayısına eşit olduğunun farkındadır.
2 puan Çok Yönlü Yapı	Öğrenci oluşturulabilecek birden fazla sayıda B kümesi gösterebilir. Yani öğrenci; {m, n}, {m, n, 1}, {m, n, 2}, {m, n, 1, 2} kümelerinden en az iki tanesini gösterir. Ancak, oluşturulabilecek B kümelerinin sayısı ile A kümesinin eleman sayısı arasındaki ilişkiyi açıklayamaz.
1 puan Tek Yönlü Yapı	Öğrenci {m, n}, {m, n, 1}, {m, n, 2}, {m, n, 1, 2} gibi oluşturulabilecek farklı B kümelerinden yalnızca birini gösterir.
0 puan Yapı Öncesi	Öğrenci probleme herhangi bir cevap vermeyip boş bırakır ya da tamamen yanlış cevap verir. A kümesinin elemanları rakamlardan ve A U B kümesinin elemanları ise, rakamlar ve harflerden oluşmaktadır gibi problemin çözüme yönelik herhangi bir bilgi sağlamayan ifadelere yer verir.

Soru 5:

Puanlama Ölçütleri	
3 puan İlişkisel Yapı	Öğrenci, Kenan'ın çiftliğinden seçilen hayvanın koyun olma olasılığı (1/4) ile İsmail'in çiftliğinden seçilen hayvanın koyun olması olasılığını (1/3) bulabilir. Ayrıca bulduğu bu iki olasılığı birbiri ile çarparak (1/4) × (1/3) = (1/12) şeklinde tutarlı bir sonuca olabilir. Hem Kenan'ın hem de İsmail'in çiftliğinden seçilen hayvanın koyun olma olasılığını doğru bir biçimde cevaplar.
2 puan Çok Yönlü Yapı	Öğrenci, Kenan'ın çiftliğinden seçilen hayvanın koyun olma olasılığı (1/4) ile İsmail'in çiftliğinden seçilen hayvanın koyun olması olasılığını (1/3) bulabilir. Ancak, bulduğu bu iki olasılıktan tutarlı bir bütün oluşturamadığından hem Kenan'ın hem de İsmail'in çiftliğinden seçilen hayvanın koyun olması olasılığını hesaplayamaz.
1 puan Tek Yönlü Yapı	Öğrenci, Kenan'ın ya da İsmail'in çiftliğinden seçilen hayvanın koyun olması olasılığını bulabilir. Yani öğrenci 1/4 veya 1/3 olasılıklarından yalnızca birini hesaplayabilir, ancak aynı anda iki olasılığı bulamaz.
0 puan Yapı Öncesi	Öğrenci, "her iki çiftlikte de hem koyun hem inek vardır" gibi cevapla ilgisi olmayan ifadeler kullanır. Öğrencinin verdiği cevap tamamıyla yanlıştır ve soruda istenilenlerle bir ilgisi yoktur.

Soru 6:

Puanlama Ölçütleri	
3 puan İlişkisel Yapı	Öğrenci, a=4 ve b=4 durumunda a=b'dir. b < a şartının sağlanması için b < 4 olmalıdır şeklinde bir genellemeye varabilir.
2 puan Çok Yönlü Yapı	Öğrenci a+b=8 ve b<a koşulunu sağlayan birden fazla b değeri olduğunun farkındadır. a=7 ve b=1, a=6 ve b=2 veya a=5 ve b=3 gibi birden fazla sayıda olası durumu yazabilir. Ancak a=4 ve b=4 durumunda a=b olduğunun b<a şartının sağlanması için b<4 olmalıdır şeklinde bir genellemeye varamaz.
1 puan Tek Yönlü Yapı	Öğrenci a=7 ve b=1, a=6 ve b=2 veya a=5 ve b=3 gibi olası durumlardan sadece birini yazabilir. b'ye birden fazla sayıda farklı değer vermeyi düşünemez.
0 puan Yapı Öncesi	Öğrenci b<8 gibi tamamıyla yanlış cevaplar verir veya bilmiyorum gibi ifadeler kullanır.

Soru 7:

Puanlama Ölçütleri	
4 puan Soyutlanmış Yapı	Öğrenci A makinesine giren sayı ile B makinesinden çıkan sayı arasındaki ilişkiyi cebirsel olarak ifade etmekle kalmaz. Soruda verilenlerin ötesinde tahminler yapıp hipotezler kurarak A makinesine giren sayı ile B makinesinden çıkan sayı arasındaki eşitlik değişmeyecek şekilde, A ve B makineleri için başka bir çalışma prensibi belirleyebilir. Öğrenci a, b, c ve d seçeneklerinin tümünü doğru yanıtlamıştır.
3 puan İlişkisel Yapı	Öğrenci, x ve y gibi bilinmeyenler arasındaki ilişkileri açıklayabilir. A makinesi ile B makinesi arasındaki ilişkiyi cebirsel olarak ifade edebilir. Öğrenci a, b ve c seçeneklerini doğru bir şekilde yanıtlamıştır. Ancak d seçeneğine cevap verememiştir.
2 puan Çok Yönlü Yapı	Öğrenci algoritmaların ve yöntemlerin uygulamasını gerektiren birden fazla işlemi yapabilir. Örneğin, A makinesine konulan farklı sayılar için B makinesinden çıkacak sayıları hesaplayabilir. Ancak A makinesine giren sayı ile B makinesinden çıkan sayı arasındaki ilişkiyi cebirsel olarak ifade edemez. Öğrenci a ve b seçeneklerini doğru cevaplayabilmiştir. Ancak c ve d seçeneklerine cevap verememiştir.
1 puan Tek Yönlü Yapı	Öğrenci soruda verilenlerin sadece bir yönüne odaklanır. Tek bir adımı gerektiren basit işlemleri çözer. A makinesi konulan bir sayı için B makinesine girecek sayının kaç olduğunu hesaplayabilir. Dolayısıyla öğrenci a seçeneğine doğru cevap verebilse de; diğer seçenekleri cevaplayamaz.
0 puan Yapı Öncesi	Öğrenci, **“**Bilmiyorum”, “Çok zor bir soru” gibi problemin çözümüne yönelik olmayan ifadeler kullanır. Ya da problemde verilenleri tekrar etmekle yetinir.

Soru 8:

Puanlama Ölçütleri	
4 puan Soyutlanmış Yapı	Öğrenci kürdan sayısı ile ev sayısı arasındaki ilişkiyi cebirsel olarak ifade etmekle kalmaz. Yan yana evler yapacak şekilde, farklı bir geometrik şekil belirleyip bu şekil için kürdan sayısı ile ev sayısı arasındaki ilişkiyi cebirsel olarak ifade edebilir. Yani öğrenci, mevcut bilgilerin ötesinde genellemelere ulaşabilir. Öğrenci a, b, c ve d seçeneklerinin tümünü doğru yanıtlamıştır.
3 puan İlişkisel Yapı	Öğrenci kürdan sayısı ile ev sayısı arasındaki ilişkiyi cebirsel olarak ifade ederek, bu ilişki hakkında bir genellemeye varabilir. Öğrenci a, b ve c seçenekleri doğru bir şekilde yanıtlamıştır. Ancak d seçeneğine cevap verememiştir.
2 puan Çok Yönlü Yapı	Öğrenci örüntünün ortak farkını kullanarak bir sonraki adımı kestirebilir. 9 ev için gerekli kürdan sayılarını hesaplamakla kalmaz. Öğrenci 42 ev için gerekli kürdan sayısına, örüntünün ortak farkını ekleyerek 43 ev için gerekli kürdan sayısını bulabilir. Ancak, kürdan sayısı ile ev sayı arasında nasıl bir ilişki olduğuna dair bir genellemeyi düşünemez ve bu ilişkiyi cebirsel olarak ifade edemez. Öğrenci a ve b seçeneklerini doğru cevaplayabilmiştir. Ancak c ve d seçeneklerine cevap verememiştir.
1 puan Tek Yönlü Yapı	Öğrenci 9 ev için gerekli kürdan sayısını soruda somut olarak verilen 3 evin yanına 6 ev daha çizerek hesaplayabilir. Ancak, 9 ev için gerekli kürdan sayısını hesaplamak için ardışık iki terim arasındaki farktan (1 ev için 5 kürdan, 2 ev için 9 kürdan, 3 ev için 13 kürdan… gibi her bir ev için kürdan sayısı 4 artar) yararlanmayı düşünemez. Bu nedenle, 9 ev için gerekli kürdan sayısını hesaplayabilse de; 42 ev için gerekli kürdan sayısından yararlanarak 43 ev için gerekli kürdan sayısını bulamaz. Dolayısıyla öğrenci a seçeneğine doğru cevap verebilse de diğer seçenekleri cevaplayamaz.
0 puan Yapı Öncesi	Öğrenci "Ben de böyle bir oyun oynamak isterdim" gibi cevapla ilgisi olmayan ifadeler kullanır.

Ek-4: Standart ve SOLO Taksonomisine Dayalı Rubrikler Hakkında Uzman Görüşü Almak için Yararlanılan Form

Sayın Uzman,
Aşağıda sekiz açık uçlu matematik sorusu ve bu soruların puanlanmasında kullanılmak üzere hazırlanan rubrikler yer almaktadır. Bu rubrikleri, tabloda verilen ölçütler doğrultusunda değerlendirmeniz rica edilmektedir. Gözden geçirilmesi gerektiğini düşündüğünüz rubrikler için düzeltme önerilerinizi açıklama olarak eklemeniz, araştırmacının uzman görüşlerini daha sağlıklı bir biçimde yorumlamasını sağlayacak ve gerekli değişiklikleri yapmasını kolaylaştırıcaktır. İlginize şimdiden teşekkür eder, çalışmalarınızda kolaylıklar dilerim…

Arş. Gör. Mustafa İLHAN

	Evet	**Kısmen**	**Hayır**
Puanlama kategorileri iyi tanımlanmış mı?			
Puan kategorileri arasındaki farklar açık mı?			
Rubrik, her nitelikteki öğrenci grubunu değerlendirmek için uygun mu?			
Rubrikte yer alan anlatımlar açık (anlaşılır) mı?			
Puanlama ölçütleri soru ile ölçülmek istenen özelliğin bütün yönlerini içeriyor mu?			
Puanlama ölçütleri soru ile ölçülmek istenen özellik dışında herhangi bir değerlendirme ölçütü içeriyor mu?			

Ek-5: Standart Rubrikler ile İlgili Düşünceler Anketi

- Standart rubriklerin objektifliği hakkında ne düşünüyorsunuz?
  ……
  ……
  ……
  ……

- Standart rubriklerin kullanım kolaylığı hakkında ne düşünüyorsunuz?
  ……
  ……
  ……
  ……

- Öğrencilere güçlü ve zayıf olduğu noktalar hakkında geri bildim vermede standart rubriklerin etkililiği hakkında ne düşünüyorsunuz?
  ……
  ……
  ……
  ……

- Standart rubrik kullanılmadan önce bu rubriklere yönelik bir eğitim verilmesinin gerekliliği hakkında ne düşünüyorsunuz?
  ……
  ……
  ……
  ……....

- Öğrencinin cevabına karşılık gelen derecenin (puanının) doğru bir biçimde belirlenmesi konusunda standart rubriklerin etkililiği hakkında ne düşünüyorsunuz?
  ……
  ……
  ……
  ……

- Ölçülen özellik açısından farklı yetenek düzeyinde bulunan öğrencilerin ayırt edebilmesi konusunda standart rubriklerin etkililiği hakkında ne düşünüyorsunuz?
  ……
  ……
  ……
  ……

- Standart rubriklere ilişkin eklemek istediğiniz görüşleriniz varsa lütfen belirtiniz.
  ……
  ……
  ……
  ……

Standart Rubriklerin;	***Çok Yüksek (5) ⟶ Çok Düşük (1)***				
Objektifliği (Puanlama işlemi kim tarafından yapılırsa yapılsın aynı sonuca ulaşılması)	5	4	3	2	1
Ölçülen özellik açısından farklı seviyelerdeki öğrencileri birbirinden ayırt edebilme özelliği	5	4	3	2	1
Öğrenciye güçlü ve zayıf olduğu noktalar hakkında geri bildirimler sunabilme özelliği	5	4	3	2	1
Standart Rubriklerin;	***Çok Kolay (5) ⟶ Çok Zor (1)***				
Hazırlanması	5	4	3	2	1
Kullanımı	5	4	3	2	1

Ek-6: SOLO Taksonomisine Dayalı Rubrikler ile İlgili Düşünceler Anketi

- SOLO taksonomisine dayalı rubriklerin objektifliği hakkında ne düşünüyorsunuz?
  ……
  ……
  ……
  ……

- SOLO taksonomisine dayalı rubriklerin kullanım kolaylığı hakkında ne düşünüyorsunuz?
  ……
  ……
  ……
  ……

- Öğrencilere güçlü ve zayıf olduğu noktalar hakkında geri bildim vermede SOLO taksonomisine dayalı rubriklerin etkililiği hakkında ne düşünüyorsunuz?
  ……
  ……
  ……
  ……

- SOLO taksonomisine dayalı rubrik kullanılmadan önce bu rubriklere yönelik bir eğitim verilmesinin gerekliliği hakkında ne düşünüyorsunuz?
  ……
  ……
  ……
  ……

- Öğrencinin cevabına karşılık gelen derecenin (puanının) doğru bir biçimde belirlenmesi konusunda SOLO taksonomisine dayalı rubriklerin etkililiği hakkında ne düşünüyorsunuz?
  ……
  ……
  ……
  ……

- Ölçülen özellik açısından farklı yetenek düzeyinde bulunan öğrencilerin ayırt edebilmesi konusunda SOLO taksonomisine dayalı rubriklerin etkililiği hakkında ne düşünüyorsunuz?
  ……
  ……
  ……
  ……

- Açık uçlu matematik soruların puanlanmasında standart ve SOLO taksonomisine dayalı rubriklerden hangisini tercih edersiniz. Neden?

  ……
  ……
  ……
  ……

- SOLO taksonomisine dayalı rubriklere ilişkin eklemek istediğiniz görüşleriniz varsa lütfen belirtiniz.

  ……
  ……
  ……
  ……

SOLO Taksonomisine Dayalı Rubriklerin;	***Çok Yüksek (5)⟶ Çok Düşük (1)***				
Objektifliği (Puanlama işlemi kim tarafından yapılırsa yapılsın aynı sonuca ulaşılması)	5	4	3	2	1
Ölçülen özellik açısından farklı seviyelerdeki öğrencileri birbirinden ayırt edilebilme özelliği	5	4	3	2	1
Öğrenciye güçlü ve zayıf olduğu noktalar hakkında geri bildirimler sunabilme özelliği	5	4	3	2	1
SOLO Taksonomisine Dayalı Rubriklerin;	***Çok Kolay (5) ⟶ Çok Zor (1)***				
Hazırlanması	5	4	3	2	1
Kullanımı	5	4	3	2	1

Ek-7: ÖZGEÇMİŞ (VITAE)

ÖZGEÇMİŞ

Mustafa İLHAN 1987 yılında Diyarbakır'da doğdu. İlköğretime Adana Yeşilyuva İlköğretim okulunda başladı ve Diyarbakır Fevzi Çakmak İlköğretim okulunda devam etti. 2001 yılında girdiği Diyarbakır Anadolu Lisesi'nden 2005 yılında mezun oldu. Aynı yıl yerleştiği Dicle Üniversitesi Ziya Gökalp Eğitim Fakültesi İlköğretim Matematik Öğretmenliği programından 2009 yılında bölüm birincisi olarak mezun oldu. 2009-2010 Öğretim Yılı'nda Diyarbakır ili Dicle İlçesi'ndeki bir ilköğretim okuluna matematik öğretmeni olarak atandı. 2009 Eylül ile 2010 Şubat tarihleri arasında bu okulda matematik öğretmeni olarak görev yaptı. 2010 yılı Şubat Ayı'nda, Dicle Üniversitesi Ziya Gökalp Eğitim Fakültesi İlköğretim Bölümü'nde araştırma görevlisi olarak çalışmaya başladı. Yazar, yüksek lisans derecesini Dicle Üniversitesi Sosyal Bilimler Enstitüsü Eğitim Bilimleri Anabilim Dalı'ndan 2011 yılında; doktora derecesini ise Gaziantep Üniversitesi Eğitim Bilimleri Enstitüsü Eğitim Bilimleri Anabilim Dalı'ndan 2015 yılında aldı. Orta düzeyde İngilizce bilen yazarın ulusal ve uluslararası dergilerde yayınlanmış makaleleri ile yurt içi ve yurt dışı bilimsel toplantılarda sunulmuş bildirileri bulunmaktadır.

VITAE

Mustafa Ilhan was born in Diyarbakir in 1987. He started elementary education at Adana Yesilyuva Elementary School, but it was at Diyarbakir Fevzi Cakmak Elementary School that he finished elementary education. He started studying at Diyarbakir Anatolian High School in 2001, from which he graduated in 2005. In the same year, he was entitled to study Elementary Mathematics Education in the Ziya Gokalp Faculty of Education at Dicle University, and he completed his studies in 2009 as the top scoring student of the department. In the 2009-2010 academic year, he was appointed as a mathematics teacher to an elementary school in Dicle, Diyarbakir. He served as a mathematics teacher at this school from September 2009 until February 2010, when he started working as a research assistant at the Department of Elementary Mathematics Teacher Education in the Ziya Gokalp Faculty of Education at Dicle University. The author had his master's degree in Educational Sciences at the Institute of Social Sciences at Dicle University in 2011, and he received his doctorate in Educational Sciences at the Institute of Educational Sciences at Gaziantep University in 2015. He has an intermediate command of English. He has published articles for national and international journals and presented papers at national and international scientific conferences.