Eğitim ve Bilim

Cilt 42 (2017) Sayı 189 217-247

Standart ve SOLO Taksonomisine Dayalı Rubrikler ile Puanlanan Açık Uçlu Matematik Sorularında Puanlayıcı Katılığı ve Cömertliğinin İncelenmesi *

Bayram Çetin [1], Mustafa İlhan [2]

Öz

Bu araştırmada, standart ve SOLO taksonomisine dayalı rubrikler ile puanlanan açık uçlu matematik sorularında puanlayıcı katılığı ve cömertliğinin çok yüzeyli Rasch modeli ile incelenmesi amaçlanmıştır. Çalışmanın veri kaynağını, sekizinci sınıfa devam eden 104 öğrencinin araştırmacılar tarafından geliştirilen matematik başarı testindeki açık uçlu sorulara verdiği cevaplar oluşturmuştur. Araştırmada puanlayıcı olarak görev alan yedi matematik öğretmeni ise çalışmanın katılımcıları olarak belirlenmiştir. Araştırma verilerinin toplanmasında, standart ve SOLO taksonomisine dayalı rubrikler kullanılmıştır. Verilerin toplanması birkaç aşamada gerçekleşmiştir. İlk aşamada açık uçlu sorulardan oluşan matematik başarı testi öğrencilere uygulanarak, puanlayıcıların değerlendirme yapacakları dokümanlar elde edilmiştir. Daha sonra puanlayıcılar, öğrencilerin açık uçlu matematik sorularına verdikleri cevapları standart rubrik kullanarak puanlamışlardır. Bu işlemin ardından SOLO taksonomisine dayalı rubrikler kullanılarak yapılan puanlamalara geçilmiştir. Açık uçlu matematik sorularına verilen cevapların standart ve SOLO taksonomisine dayalı rubrikler kullanılarak puanlanmasıyla elde edilen veriler çok yüzeyli Rasch modeline göre analiz edilmiştir. Araştırmada; standart rubrikler kullanılarak yapılan puanlamalarda; puanlayıcılar arası uyumun düşük olduğu ve katılık/cömertlikleri yönüyle puanlayıcılar arasında anlamlı fark bulunduğu belirlenmiştir. SOLO taksonomisine dayalı rubriklerden yararlanılarak yapılan puanlamalarda ise puanlayıcılar arası uyumun yüksek olduğu ve puanlayıcıların benzer katılık/cömertlikte puanlama yaptıkları saptanmıştır.

Anahtar Kelimeler

Açık uçlu sorular
Puanlayıcı katılığı ve cömertliği
Rubrik
SOLO taksonomisi
Çok yüzeyli Rasch modeli

Makale Hakkında

Gönderim Tarihi: 29.07.2015
Kabul Tarihi: 07.11.2016
Elektronik Yayın Tarihi: 21.02.2017

DOI: 10.15390/EB.2017.5082

* Bu çalışma, Mustafa İlhan'ın doktora tezinden türetilmiştir.

[1] Gazi Üniversitesi, Gazi Eğitim Fakültesi, Eğitim Bilimleri Bölümü, Türkiye, bcetin27@gmail.com
[2] Dicle Üniversitesi, Ziya Gökalp Eğitim Fakültesi, İlköğretim Bölümü, Türkiye, mustafailhan21@gmail.com

Giriş

Matematik Başarısının Değerlendirilmesi

Başarı doğrudan gözlenemeyip ancak dolaylı olarak ölçülebilen soyut bir yapı olduğundan (Tan, 2015), öğrencilerin herhangi bir matematik konusunu ne düzeyde öğrendiğinin belirlenmesi oldukça zordur. Öğrencinin konuyu ne düzeyde öğrendiğine dair yapılan çıkarımlarda, bir göreve ilişkin ortaya koyduğu performans ya da kendisine yöneltilen 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 cevapların, 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)* düşüncelerini açıklamak için matematiksel terminolojiyi doğru bir şekilde kullanma, *iv)* tümevarım ve tümdengelim düşünce süreçlerini kullanarak çıkarımlar yapma, *v)* problem çözme stratejileri geliştirme ve bunları günlük hayattaki problemlere uygulama gibi matematik eğitimi kapsamında öğrencilere kazandırılması öngörülen becerilerin (Milli Eğitim Bakanlığı [MEB], 2009; Matematik Öğretmenleri Ulusal Konseyi [NCTM], 2000) öğrenciler tarafından ne düzeyde kazanıldığını ortaya çıkarabilecek nitelikte olmalıdır. Çoktan seçmeli testler sıralanan becerilerin ölçülmesinde yetersiz kaldığından, matematik değerlendirme sürecinde çoktan seçmeli testlerin ötesinde yöntemlere ihtiyaç duyulmaktadır. Bu ihtiyacı performans değerlendirmenin doğasında var olan özellikler karşılayabilmektedir (Güler, 2008).

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 (Stecher, 2010). Performans değerlendirmede, öğrencinin sunulan seçeneklerden herhangi birini seçmesi yerine; cevabı 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). Diğer bir deyişle performans değerlendirme; öğrencinin karmaşık problemleri çözmesini, problemi çözmek için kullandığı süreçleri göstermesini (McBee ve Barnes, 2009) ve cevabının gerekçelerini açıklamasını gerektirmektedir (Woodward, Monroe ve Baxter, 2001). Performans değerlendirmeye ilişkin bu özellikler, öğrencinin güçlü ve zayıf yönlerini görebilmesini, öğrendiklerine ilişkin daha ayrıntılı bilgi sahibi olmasını, 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; üst düzey zihinsel becerilerinin gelişmesine yardımcı olmaktadır (Kind, 1999; National Assessment Governing Board [NAGB], 2002). 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 yukarıda sıralanan avantajlarının yanında birtakım sınırlılıkları bulunmaktadır. Bu sınırlılıklardan en önemlisi, performans değerlendirmenin çoktan seçmeli testler gibi objektif bir biçimde puanlanamamasıdır (Romagnano, 2001). Öğrencilerin objektif olarak puanlanamayan herhangi bir testten aldığı puan, testi puanlayan kişiye göre farklılık göstermektedir (Tekin, 2009). Alanyazında bu durumu örneklendiren çalışmalar (Özmantar, Bingölbali ve Akkoç, 2008; Güler, 2008; Kan, 2005; Koretz, McCaffrey, Klein, Bell ve Stecher, 1992; Toffoli, Andrade ve Bornia, 2016) bulunmaktadır. Örneğin, Koretz ve diğerlerinin (1992) yaptığı çalışmada öğrencilerin matematik performansları dörtlü derecelemeye sahip bir rubrik kullanılarak iki puanlayıcı tarafından puanlanmış ve puanlayıcılar arası uyumun düşük olduğu tespit edilmiştir. Özmantar ve diğerleri (2008) tarafından yapılan araştırmada, 171 öğretmen açık uçlu bir matematik sorusuna verilen aynı öğrenci cevabını

puanlamış ve öğretmenlerin aynı cevaba 0 ile 10 arasında yer alan geniş bir yelpazede oldukça farklı notlar verdiği görülmüştür. Aynı cevaba yönelik olarak, öğretmenlerin %44'ü 10 üzerinden 10 tam puan verirken; %24'ü 0 puan vermiştir. Bingölbali, Özmantar ve Akkoç (2008) tarafından yapılan bir başka araştırmada, öğrencilerin açık uçlu matematik sorularına verdikleri cevapları 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 cevaplar dört farklı puanlayıcı tarafından puanlanmış ve elde edilen veriler çok yüzeyli Rasch modeline göre analiz edilmiştir. Rasch analizi sonucunda, puanlayıcılar arası uyumun düşük olduğu ve puanlayıcıların aynı cevaba farklı puanlar verme eğiliminde oldukları belirlenmiştir. Bu sonuçlar, açık uçlu sorularda öğrencinin performansının yalnızca yetenek düzeyine bağlı olmadığını; puanlayıcı kaynaklı faktörlerden de (puanlayıcının yaşı, cinsiyeti, puanlama tecrübesi, daha önce aldığı puanlama eğitimleri vb.) etkilendiğini göstermektedir. Öğrencinin performansını etkileyen puanlayıcı kaynaklı faktörler *puanlayıcı etkisi* olarak adlandırılmaktadır (Farrokhi, Esfandiari ve Vaez Dalili, 2011). Puanlayıcı etkileri, öğrencinin test puanlarına ölçülen yapı ile ilgisi olmayan varyansın karışmasına neden olduğundan (Eckes, 2005; Hoyt, 2000), ölçme işlemine karışan hatayı arttırmakta ve öğrencinin yetenek düzeyi hakkında verilen kararın güvenirliğini düşürmektedir.

Performans değerlendirmeye karışan puanlayıcı etkileri; puanlayıcı katılığı ve cömertliği, merkeze yönelme etkisi, halo etkisi, ranj sınırlaması (Saal, Downey ve Lahey, 1980), yanlılık ve tutarsızlık gibi farklı başlıklar altında incelenmektedir (Myford ve Wolfe, 2004). Bununla birlikte; Cronbach (1990) puanlayıcı katılığı ve cömertliğini puanlama sürecine karışan en önemli puanlayıcı etkisi olarak nitelendirmiştir. 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). Bu eğilim puanlayıcılar arasındaki tutarlılığı azaltmaktadır. Aynı bireye iki ya da daha fazla sayıda farklı puanlayıcı tarafından verilen puanlar arasındaki tutarlılık, puanlayıcılar arası güvenirliğin ölçüsü olarak alınmaktadır (Moskal ve Leydens, 2000). Dolayısıyla birden fazla sayıda puanlayıcının farklı katılık ve cömertlikte puanlamalar yapması puanlayıcılar arası güvenirliğin düşük; benzer katılık ve cömertlikte puanlamalar yapması ise puanlayıcılar arası güvenirliğin yüksek olduğu anlamına gelmektedir.

Puanlayıcının katılığı ve cömertliği; performansı katı bir puanlayıcı tarafından değerlendirilen bir öğrencinin, ölçülen özelik açısından kendisinden daha az yetenekli olan ancak daha cömert bir puanlayıcı tarafından değerlendirilen bir öğrenciye göre daha düşük bir puan almasına neden olabilmektedir (Wiseman, 2012). 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 (yetenek düzeyi değerlendirmede kullanılan ölçütün bulunduğu noktaya karşılık gelen) öğrenciler için telafisi zor olan 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 cömertliği ise, öğrenme hedeflerine ulaşamayan bir öğrencinin dersi geçmesine veya okuldan mezun olmasına sebebiyet verebilir. Söz gelimi, tıp fakültesinde öğrenim gören öğrencilerin dikiş atma performanslarının cömert bir puanlayıcı tarafından değerlendirilmesi halinde, dikiş atma becerisi ile ilgili önemli eksiklikleri bulunan bir öğrencinin ilgili kazanıma ulaştığı şeklinde yanlış bir karar alınabilir. Dolayısıyla, öğrencinin yetenek düzeyine ilişkin doğru kestirimler elde edilebilmesi için puanlayıcı katılığı ve cömertliğinin minimum düzeyde tutulması gerekir. Puanlayıcı katılığı ve cömertliğini kontrol altına alabilmek için değerlendirmede kullanılacak ölçütlere ilişkin puanlayıcılar arasında ortak bir anlayış oluşturulmalıdır. Bu ortak anlayışın oluşmasına hizmet edecek en önemli uygulamalardan biri, puanlamaların rubriklere dayalı olarak yapılmasıdır.

Standart ve SOLO (Structure of Observed Learning Outcome) Taksonomisine Dayalı Rubrikler

Rubrikler, performansın değişik düzeylerine ait karakteristik özellikler ile ölçütleri tanımlayan ve bu özellik ile ölçütler doğrultusunda performansa ilişkin yargıya varmada kullanılan puanlama rehberidir (Kan, 2007). Rubrikler, bir yandan aynı performansı değerlendiren farklı puanlayıcıların verdiği puanlar arasındaki tutarlılığı arttırırken; diğer yandan bir puanlayıcının aynı performansa farklı

zamanlarda değişik puanlar vermesinin önüne geçmektedir. Rubrikler bu sayede, puanlama işleminin ne zaman ve kim tarafından yapıldığından bağımsız olarak gerçekleştirilmesini sağlar (Moskal ve Leydens, 2000). Rubrikler bazen herhangi bir taksonomi temele alınmadan geliştirilmektedir. Herhangi bir taksonomi temele alınmadan geliştirilen rubrikler için bu çalışmada *standart rubrik* ifadesi kullanılmıştır. 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 (Gronlund, 1998). 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; standart bir rubrikte açık uçlu bir matematik problemi için öğrencinin hem ulaştığı sonucun hem de çözüm yolunun hatalı olduğu cevaplar *yetersiz* düzeyine; çözüm yolunun doğru ancak yapılan işlemlerin ve bulunan sonucun hatalı olduğu cevaplar *geliştirilmesi gerek* düzeyine; küçük işlem hataları nedeniyle doğru sonuca ulaşılamayan ancak çözümde takip edilen işlem basamaklarının doğru, açık ve anlaşılır olduğu cevaplar *iyi* düzeyine, bulunan sonucun doğru, çözümde takip edilen basamakların açık ve anlaşılır olduğu cevaplar 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 1'de sunulmuştur.

Tablo 1. Matematik Problem Çözme Becerisi İçin Standart Rubrik Örneği

Ölçüt		**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 sonuca nasıl ulaşıldığı 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	
	-Problem 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; yansıtıcı düşünme modeli, Bloom taksonomisi ya da SOLO taksonomisi gibi farklı modellerden de yararlanılabilmektedir (Chan, Tsui, Mandy ve Hong, 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 Biggs ve Collis (1982) tarafından ileri sürülmüştür. SOLO taksonomisine göre öğrenme döngüsü; yapı öncesi, tek yönlü yapı, çok yönlü yapı, ilişkisel yapı ve soyutlanmış yapı olmak üzere beş düzeyli bir yapıya sahiptir (Mohd Nor ve Idris, 2010). Bu beş düzey, öğrencinin herhangi bir soruyu yanıtlarken cevabını beş farklı şekilde yapılandırabileceğini göstermektedir (Lucas ve Mladenovic, 2008). Yapı öncesi düzeyde, öğrenci gerçekleştirmesi beklenen görevi uygun bir biçimde yerine getirememektedir Öğrencinin ileri sürdüğü fikirler problemin çözümü için herhangi bir yarar sağlamamaktadır (Leung, 2000). Öğrencinin verdiği cevabın çözmeye çalıştığı problemle çok fazla ilgisi yoktur (Brabrand ve Dahl, 2009). Tek yönlü yapı düzeyinde, öğrenci konuyu dar ve yüzeysel bir bakış açısı ile ele almakta ve konunun tek bir yönüne odaklanmaktadır. Çok yönlü yapı düzeyinde, öğrenci konu ile ilgili birden fazla yönü anlamakta, ancak bunlar arasında bir ilişki kuramamaktadır. 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 (Leung, 2000). İlişkisel yapı düzeyinde öğrenci olayın çeşitli yönlerini görüp bunlardan anlamlı bir bütün oluşturabilmektedir. Bu düzeyde kavramlar benzer bir duruma ya da probleme uygulanabilmekte; ancak farklı bir alana transfer edilememektedir (Kanuka, 2011). Soyutlanmış yapı ise öğrencinin yansıtma ve değerlendirme yapabildiği, hipotez kurabildiği, tümevarım, tümdengelim ve kombinasyonel düşünme süreçlerini kullanarak öğrendiklerini farklı bir alana transfer edebildiği bir düzey olarak tanımlanmaktadır (Lake, 1999).

SOLO Taksonomisine Dayalı Rubriklerin Avantajları

Öğrenci performansının değerlendirilmesi sürecinde SOLO taksonomisine dayalı rubrik kullanımının birçok avantajı bulunmaktadır. Öncelikle, SOLO taksonomisi öğrencilerin öğrenme sürecindeki eksikliklerini görmeye yardımcı olmakta ve değerlendirme sürecinde kısmi puanlamalar yapılmasına imkân tanımaktadır. Bu özelliği, SOLO taksonomisini düzey belirlemeye yönelik değerlendirmelerin yanı sıra öğrencilerin güçlü ve zayıf yönlerinin belirlenip, eksiklik ve hatalarının düzeltilmesi amacı ile gerçekleştirilen biçimlendirici değerlendirmeler için de uygun bir araç haline getirmektedir (Hattie ve Purdie, 1998). Öğrenmelerin hem niteliksel (derin öğrenme) hem de niceliksel (yüzeysel öğrenme) yönünün belirlenebilmesine imkân tanıması SOLO taksonomisine dayalı rubriklerin bir diğer güçlü yönüdür (Burnett, 1999). SOLO taksonominin ilk basamağı olan yapı öncesi düzeyde öğrencinin konuya ilişkin herhangi bir öğrenmesi yoktur veya öğrenmeleri tamamıyla yanlıştır. Tek yönlü yapı düzeyinde, öğrencinin konu hakkında bildikleri konunun yalnızca bir yönü ile sınırlıdır. Çok yönlü yapı düzeyinde ise öğrenci konu ile ilgili birkaç fikri sıralayabilmekte fakat bu fikirleri birbirleriyle ilişkilendirip tutarlı bir bütüne ulaşamamaktadır. Bu anlamda yapı öncesi ile çok yönlü yapı düzeyleri arasında öğrencinin konuya ilişkin öğrenmelerinde yalnızca niceliksel bir artış meydana gelmektedir (Rembach ve Dison, 2016). Öte yandan, ilişkisel yapı düzeyinde öğrenci konu hakkında öğrendiklerini birbirileri ile ilişkilendirerek tutarlı bir bütüne ulaşabilmekte ve soyutlanmış yapı düzeyinde ilişkisel yapı basamağında oluşturduğu anlamlı bütünü daha farklı bir bağlama/konuya genelleyebilmektedir. Dolayısıyla, ilişkisel ve soyutlanmış yapı basamakları öğrenmedeki niteliksel dönüşümü yansıtmaktadır (Brabrand ve Dahl, 2009). Sıralanan avantajlarından dolayı, birçok farklı derste açık uçlu soruları puanlamak üzere SOLO taksonomine dayalı rubriklerden yararlanılmaktadır. Bu derslerden biri de matematiktir (Collis ve Romberg, 1992; Lian ve Yew, 2012). SOLO taksonomisi doğrudan matematik dersi öğrenme çıktılarının değerlendirilmesi amacıyla geliştirilmiş bir teori olmasa da; taksonominin düzeyleri cebir, istatistik ve geometri gibi matematiksel düşünmenin farklı biçimleri ile paralellik göstermektedir (Jurdak, 1991; Lian ve Idris, 2006; Mooney, 2002). Bundan dolayı, SOLO taksonomisine dayalı rubrikler açık uçlu matematik sorularının puanlanmasında sıklıkla kullanılmaktadır. Bu yaygın kullanımına rağmen açık uçlu matematik sorularında puanlayıcı etkilerini kontrol altına alma konusunda SOLO taksonomisine dayalı rubriklerin ne derece işlevsel olduğu sorusuna cevap olabilecek bir çalışmaya alanyazında rastlanmamaktadır.

Araştırmanın Amacı ve Önemi

Bu araştırmada, standart ve SOLO taksonomisine dayalı rubrikler ile puanlanan açık uçlu matematik sorularında puanlayıcı katılığı ve cömertliğinin incelenmesi amaçlanmaktadır. Araştırma sonucunda; açık uçlu matematik sorularını puanlamak üzere geliştirilen iki rubrik türünden hangisinin puanlayıcılar arasındaki farklılıkları giderme ve puanlayıcı güvenirliğini arttırma konusunda daha etkili olduğu belirlenecektir. Buna bağlı olarak; araştırma bulgularının öğrencilerin matematik performanslarının değerlendirilmesine yönelik çalışmalara yön verici bir niteliğe sahip olacağı umulmaktadır. Ayrıca; SOLO taksonomi içerikten bağımsız bir model olduğundan (Kanuka, 2011); araştırmada ulaşılacak bulguların dolaylı olarak farklı disiplinlerdeki ölçme-değerlendirme çalışmalarına da kılavuzluk etmesi beklenmektedir. Milli Eğitim Bakanlığı (MEB) ile Öğrenci Seçme ve Yerleştirme Merkezi'nin (ÖSYM) ilerleyen yıllarda ulusal düzeydeki sınavlarda açık uçlu sorulara yer vermeyi planladığı dikkate alındığında; araştırmanın geniş ölçekli test uygulamalarına da önemli katkılar sağlayacağı tahmin edilmektedir. MEB tarafından temel eğitimden ortaöğretime geçiş sistemi ile ilgili olarak yapılan açıklamada, FATİH (Fırsatları Arttırma ve Teknolojiyi İyileştirme Hareketi) 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; ÖSYM 2013 yılı itibariyle, Açık Uçlu Sorularla Sınav Projesi'ni başlatmıştır (Öğrenci Seçme ve Yerleştirme Merkezi [ÖSYM], 2013). 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. Dolayısıyla, çalışmadan elde edilen bulguların geniş ölçekli test uygulamaları için de işlevsel bir niteliğe sahip olması beklenmektedir.

SOLO taksonomisine dayalı rubrikler ile yapılan değerlendirmelerde puanlayıcı güvenirliğinin yüksek olacağına yönelik kuramsal bilgiler alanyazında geniş bir yer tutmaktadır. Ancak bu rubriklerin puanlayıcı güvenirliği üzerindeki etkisini ampirik olarak inceleyen çalışmaların (Burnett, 1999; Chan vd., 2002; Hattie ve Purdie, 1998; Leung, 2000; Yazıcı, 2013) sayısı oldukça sınırlıdır. Sözü geçen az sayıdaki çalışmada, SOLO taksonomisine dayalı rubriklerin puanlayıcı güvenirliğini nasıl etkilediğine ilişkin çelişkili bulgular elde edilmiştir. Örneğin; Hattie ve Purdie (1998); Burnett (1999) ve Chan ve diğerleri (2002), SOLO taksonomisine dayalı rubriklerin puanlayıcı güvenirliğini arttırdığını rapor ederken; Leung (2000) ile Chan, Hong ve Chan (2001), SOLO taksonomisine dayalı rubrik kullanarak yaptıkları değerlendirmelerde, puanlayıcılar arası güvenirliğin düşük olduğu sonucuna ulaşmıştır. Araştırmalarda birbirinden farklı sonuçlar elde edilmesi, bu konuda yeni çalışmalara ihtiyaç olduğunu ortaya koymaktadır. Bu araştırmanın bahsi geçen ihtiyaca cevap olabileceği ve dolayısıyla konu ile ilgili alanyazına katkı sağlayacağı düşünülmektedir.

SOLO taksonomisine dayalı rubriklerin, puanlayıcı güvenirliğini nasıl etkilediği sorusuna isabet derecesi yüksek cevaplar verebilmek için bu konuda yapılacak araştırmaların dikkatli bir biçimde planlanması gerekmektedir. Alanyazındaki çalışmalara bakıldığında, SOLO taksonomisine dayalı rubriklerin puanlayıcı güvenirliği üzerindeki etkisinin standart rubrikler ile karşılaştırılmadan incelendiği görülmektedir. SOLO taksonomisinin puanlayıcı güvenirliğini nasıl etkilediği; Hattie ve Purdie'nin (1998) yaptığı çalışmada Bloom taksonomisiyle, Çetin, Boran ve Yazıcı (2014) tarafından yapılan çalışmada puanlayıcıların kendi hazırladıkları puanlama anahtarıyla, Chan ve diğerlerinin (2002) yaptığı araştırmada ise Bloom taksonomisi ve yansıtıcı düşünme modeline dayalı rubrikler ile karşılaştırılarak belirlenmeye çalışılmıştır. Hangi rubrik türü kullanılırsa kullanılsın, rubriklerin puanlayıcılar arası farklılıkları azaltarak puanlayıcı güvenirliğini arttırması beklenmektedir (Airasian, 2005). Dolayısıyla, SOLO taksonomisine dayalı rubriklerin puanlayıcılar arası farklılıkları giderme ve puanlayıcı güvenirliğini arttırma konusunda 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 cevaplanması 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 (Burnett, 1999; Hundzynski, 2008; Leung, 2000; Yazıcı, 2013); puanlayıcılar arası korelasyon katsayısı, basit uyum yüzdesi ve puanlayıcı ortalamalarının karşılaştırılması gibi klasik test kuramına dayalı tekniklerin kullanıldığı görülmektedir. Bu araştırmada ise öğrencilerin açık uçlu matematik sorularına verdikleri cevapların standart ve SOLO taksonomisine dayalı rubrikler ile puanlanması sonucunda elde edilecek veriler çok yüzeyli Rasch modeline göre analiz edilecektir. Çok yüzeyli Rasch modeline göre; açık uçlu sorularla ölçme yapılması durumunda öğrencilerin test puanlarını etkileyebilecek faktörler bireylerin yetenek düzeyleri ya da ölçme işleminde kullanılan maddelerin güçlük düzeyleri ile sınırlı kalmamakta; puanlayıcılar ile ilgili faktörler de öğrencinin test puanlarında değişkenliğe yol açabilmektedir (Baird, Hayes, Johnson, Johnson ve Lamprianou, 2013). Bu özelliği çok yüzeyli Rasch modelini öznel olarak puanlanan açık uçlu sorular için uygun bir seçenek haline getirmektedir (Mulqueen, Baker ve Dismukes, 2000). Çok yüzeyli Rasch modeli ayrıca; aynı anda birden fazla hata kaynağını dikkate alması, farklı hata kaynakları arasındaki etkileşimleri belirleyebilmesi (Haiyang, 2010), geçerliği daha yüksek yetenek kestirimleri üretmesi (İlhan, 2016), puanlayıcı ya da performansı değerlendirilen bireyler için grup düzeyinde bilgi vermek yerine bireysel düzeyde bilgi sunması (Barkaoui, 2008) bakımından klasik test kuramına göre psikometrik açıdan daha güçlü bir modeldir. Sıralanan gerekçelerle; SOLO taksonomisine dayalı rubriklerde puanlayıcı güvenirliğinin klasik test kuramına dayalı yöntemler ile incelendiği önceki araştırmalardan farklı olarak, bu çalışmada çok yüzeyli Rasch modeli kullanılacaktır. Araştırma bu yönüyle de ilgili alanyazına önemli bir katkı sunacaktır.

Yöntem

Bu bölümde; araştırmanın veri kaynağı ile katılımcıları, çalışmada kullanılan veri toplama araçları, araştırma süresince takip edilen işlemler ve veri analizinde kullanılan istatistiksel teknikler açıklanmıştır.

Veri Kaynağı

Araştırmanın veri kaynağını, sekizinci sınıfa devam eden 104 (46 bayan ve 58 erkek) öğrencinin araştırmacılar tarafından geliştirilen başarı testindeki açık uçlu sorulara verdikleri cevaplar oluşturmuştur. Öğrencilere uygulanan başarı testi matematik dersine yönelik olarak geliştirilmiştir. Test geliştirme sürecinde öncelikle; sayılar, geometri, cebir, ölçme, olasılık ve istatistik öğrenme alanlarına yönelik toplam 18 açık uçlu soru hazırlanmıştır. 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 2'de sunulmuştur. 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. Uzmanlardan alınan görüşler doğrultusunda anlaşılırlık açısından problem oluşturabileceği ifade edilen ya da sekizinci sınıf düzeyine uygun olmadığı düşünülen altı madde ölçme aracından çıkarılmış ve bazı maddelerin ifade ediliş şekillerinde değişikliğe gidilmiştir.

Uzman görüşlerinden sonra; altı maddenin ölçekten çıkarılması ve getirilen öneriler doğrultusunda beş maddede gerekli değişikliklerin yapılmasının ardından, 12 sorudan 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. 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 soru belirlenmiştir. Ancak esas uygulamada 10 soruluk 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 rastlanmıştır. Ön uygulamanın yapıldığı grupta testi en kısa sürede tamamlayan öğrenci ile en geç tamamlayan öğrencinin kullandıkları süreler dikkate alınarak testin uygulama süresi 40 dakika olarak belirlenmiştir.

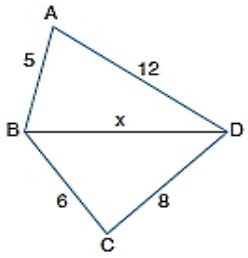
Tablo 2. Hazırlanan Açık Uçlu Soruları Anlaşılırlık ve Sekizinci Sınıf Düzeyine Uygunluk Açısından Değerlendiren Uzmanlara İlişkin Demografik Özellikler

Uzmanlar	Cinsiyet	Eğitim Durumu
1	Erkek	Matematik eğitimi alanında doçenttir.
2	Erkek	Matematik eğitimi alanında doçenttir.
3	Bayan	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	Bayan	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	Bayan	İlköğretim matematik öğretmenliği alanından mezundur. Eğitimde ölçme değerlendirme alanında yüksek lisans eğitimine devam etmektedir.
10	Bayan	İlköğretim matematik öğretmenliği alanından mezundur. Eğitimde ölçme değerlendirme alanında yüksek lisans eğitimine devam etmektedir.

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. SOLO taksonomisinin tek yönlü ile çok yönlü yapı düzeyindeki sorular hatırlama ve ezberleme düzeyindeki davranışları ölçmeye yönelik olduğundan, bu düzeylere karşılık gelen soruların çoktan seçmeli madde formatında sunulmasının açık uçlu madde formatında sunulmasına göre daha ekonomik olacağına karar verilmiştir. Bundan dolayı testte tek yönlü ve çok yönlü yapı düzeyine karşılık gelen sorulara yer verilmemiştir. Soyutlanmış yapı düzeyi; genellemelere varmayı, hipotez kurmayı, tümevarım, tümdengelim ve kombinasyonel akıl yürütme süreçlerini kullanmayı gerektirdiğinden, bu düzeye karşılık gelen soruları cevaplayabilmesi öğrencide soyut düşünme biçiminin gelişmesiyle yakından ilgilidir. Bilişsel gelişimin özellikleri dikkate alındığında, sekizinci sınıf öğrencilerinde soyut düşünmenin gelişmeye başladığı ancak tam olarak kazanılmadığı (Erden ve Akman, 2011) düşünülmüştür. Bu sebeple testte soyutlanmış yapı düzeyine karşılık gelen sorulara yer verilmiş fakat bu soruların sayısı iki ile sınırlandırılmıştır. İlişkisel yapı

düzeyine karşılık gelen sorular ise ilişkilendirme, analiz etme, sebep ve sonuçları açıklama gibi davranışları ölçmeye yönelik olduğundan, bu düzeydeki sorular sekizinci sınıf öğrencilerinin bilişsel gelişimleri için daha uygun görülmüştür. Bundan dolayı, testte soyutlanmış yapı düzeyine yönelik sorular ile kıyaslandığında, ilişkisel yapı düzeyindeki sorulara daha fazla yer verilmiştir. Matematik başarı testinden ilişkisel ve soyutlanmış yapı düzeylerine örnek olabilecek birer soru Tablo 3'te sunulmuştur.

Tablo 3. Matematik Başarı Testinden İlişkisel ve Soyutlanmış Yapı Düzeyine Yönelik Soru Örnekleri



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.

Yanda verilen soruda, öğrencinin her iki üçgende de üçgen eşitsizliğini uygulaması ve sonrasında elde ettiği eşitsizlikleri bir araya getirerek tutarlı bir bütüne ulaşması beklenmektedir. Bundan dolayı soru SOLO taksonomisinin ilişkisel yapı basamağına karşılık gelmektedir.

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.

Yandaki sorunun *a seçeneği* tek yönlü yapı düzeyinde olup üç evin yanına altı ev daha çizmesi öğrencinin bu seçeneğe doğru cevap verebilmesi için yeterlidir. Sorunun *b seçeneği* çok yönlü yapı düzeyine karşılık gelmektedir. Ev ile kürdan sayısı arasında cebirsel bir ilişki kuramayan ancak örüntünün ortak farkını hesaplayabilen bir öğrencinin bu seçeneğe doğru cevap vermesi mümkündür. Sorunun *c seçeneği* öğrencinin ev ve kürdan sayısı arasında cebirsel bir ilişki kurmasını gerektirmektedir. Ancak kurulacak ilişki soruda verilenler ile sınırlı olduğundan c seçeneği ilişkisel yapı düzeyine karşılık gelmektedir. Sorunun *d seçeneği* ise öğrencinin verilen bilgilerin ötesinde bir ilişki kurması gerektirdiğinden soyutlanmış yapı düzeyine yöneliktir. SOLO taksonomisinin hiyerarşik yapısı gereği, bir sorunun bilişsel düzeyine karar verirken soru ile ölçülmek istenen en üst basamak esas alınmaktadır. Bu açıdan yandaki soru soyutlanmış yapı düzeyine karşılık gelmektedir. Bununla birlikte, öğrencinin hangi seçeneğe kadar doğru cevap verebildiği dikkate alınarak kısmi puanlamalar yapılabilmektedir.

Katılımcılar

Araştırmanın katılımcıları, öğrenciler tarafından cevaplanan açık uçlu soruların değerlendirilmesinde görev alan yedi puanlayıcıdan (üç bayan ve dört erkek) oluşmaktadır. Puanlayıcılar araştırmaya gönüllü olarak katılmış ve kolay ulaşılabilirlik ilkesi göz önünde bulundurularak belirlenmiştir. Araştırmaya dâhil edilen puanlayıcıların demografik özelliklerine ilişkin bilgiler Tablo 4'te sunulmuştur. Tablo 4'te yer verildiği üzere; puanlayıcılardan biri matematik eğitimi alanında, diğer altısı eğitimde ölçme değerlendirme alanında yüksek lisans yapmaktadır. Puanlayıcıların eğitim durumuna ilişkin bu özelliklerin araştırma bulgularının genellenebilirliği açısından bir sakıncasının olmayacağı düşünülmektedir. Çünkü araştırmada standart ve SOLO taksonomisine dayalı rubriklerin karşılaştırılmasına odaklanılmış olup, her iki rubrik türüne göre yapılan puanlamalar aynı puanlayıcılar tarafından gerçekleştirilmiştir.

Tablo 4. Puanlayıcılara İlişkin Demografik Bilgiler

Puanlayıcı	Cinsiyet	Yaş	Öğretmenlikteki Görev Süresi	Eğitim Durumu
P1	Bayan	22	-	İlköğretim matematik öğretmenliği mezunudur ve matematik eğitimi alanında yüksek lisans yapmaktadır.
P2	Bayan	22	7 ay	İlköğretim matematik öğretmenliği mezunudur ve eğitimde ölçme değerlendirme alanında yüksek lisans yapmaktadır.
P3	Bayan	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 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.

Veri Toplama Araçları

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 deyişle, 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 1-6 numaralı sorular 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. Ek 1'de, testte yer alan açık uçlu sorulardan biri örnek olarak sunulmuş ve bu sorunun puanlanmasında kullanılan standart ve SOLO taksonomisine dayalı rubriklere yer verilmiştir.

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. Standart rubrikler hakkında görüşüne başvurulan uzmanların demografik özelliklerine ilişkin bilgiler Tablo 5'te gösterilmiştir.

Tablo 5. Standart Rubrikler Hakkında Görüşüne Başvurulan Uzmanların Demografik Özelliklerine İlişkin Bilgiler

Uzman	Cinsiyet	Eğitim Durumu
1	Erkek	Eğitimde ölçme ve değerlendirme alanında doçenttir.
2	Erkek	Sınıf öğretmenliği alanında doçenttir.
3	Bayan	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)* rubriklerdeki ifadelerin anlaşılır olduğunu, *ii)* puanlama kategorilerinin iyi tanımlandığını, *iii)* puanlama kategorileri arasındaki farkların açık olduğunu, *iv)* rubriklerin her nitelikteki öğrenci grubunu ölçmek için kullanılabileceğini, *v)* puanlama ölçütlerinin soru ile ölçülmek istenen özelliğin bütün yönlerini yansıttığını ve ölçülmek istenen özellik dışında herhangi bir değerlendirme ölçütü içermediğini göstermiştir. Dolayısıyla uzman görüşleri; standart rubriklerin kullanıma hazır olduğu ve herhangi bir değişikliğe ihtiyaç duyulmadığı şeklinde değerlendirilmiştir.

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 cevap 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. Görüşüne başvurulan uzmanların demografik özelliklerine ilişkin bilgiler Tablo 6'da verilmiştir.

Tablo 6. SOLO Taksonomisine Dayalı Rubrikler Hakkında Görüşüne Başvurulan Uzmanların Demografik Özelliklerine İlişkin Bilgiler

Uzman	Cinsiyet	Eğitim Durumu
1	Erkek	Eğitimde ölçme ve değerlendirme alanında doçenttir.
2	Bayan	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 eğitime devam etmektedir ve matematik öğretmeni olarak görev yapmaktadır.

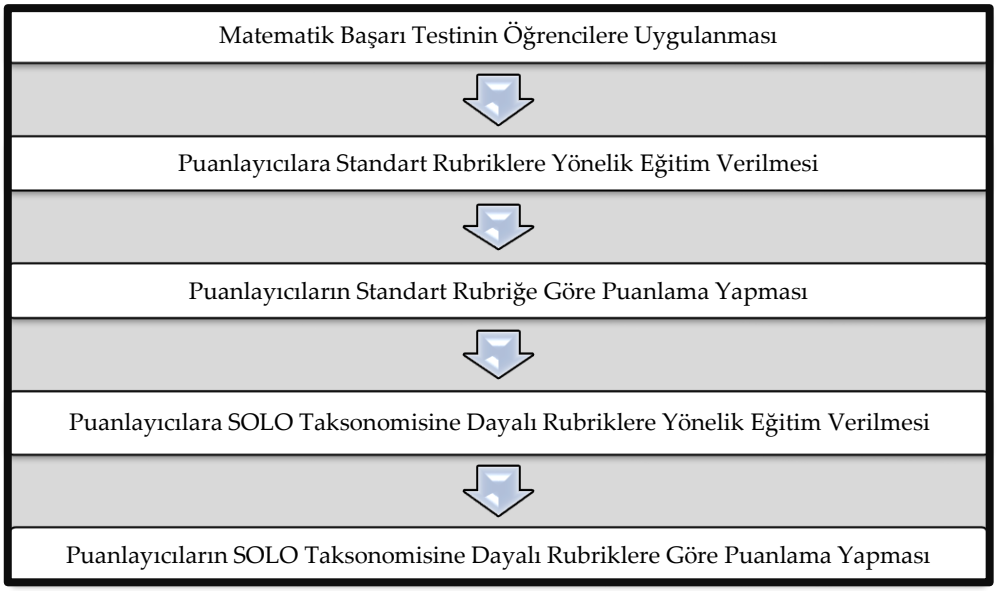
Uzmanlardan alınan görüşler; *i)* rubriklerdeki ifadelerin anlaşılır olduğunu *ii)* puanlama kategorilerinin SOLO taksonomisinin düzeylerine uygun olarak tanımlandığını, *iii)* puanlama kategorileri arasındaki farkların açık olduğunu, *iv)* rubriklerin her nitelikteki öğrenci grubunu ölçmek için kullanılabileceğini,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

İşlem

Araştırmanın verileri, 2013-2014 Öğretim Yılı Bahar Dönemi'nde toplanmıştır. İlk aşamada, puanlayıcıların değerlendirme yapacakları dokümanları elde edebilmek için araştırmacılar tarafından geliştirilen matematik başarı testi, sınıf ortamında öğrencilere uygulanmıştır. Öğrencilerden soruların çözümüne yönelik yaptıkları işlemleri açık bir biçimde yazmaları istenmiştir. Başarı testi öğrencilere uygulandıktan sonra, kâğıtlar numaralandırılmış ve 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. Böylece puanlayıcıların değerlendirme yapacakları dokümanlara ulaşılmıştır. Standart ve SOLO taksonomisine dayalı rubriklere göre yapılan puanlamalar, araştırmanın veri kaynağı olarak isimlendirilen bu dokümanlar üzerinde gerçekleştirilmiştir.

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, değerlendirmelerin yapılacağı dokümanlar elde edildikten sonra puanlayıcılara standart rubriklerin tanıtıldığı ve puanlayıcıların standart rubrik kullanarak örnek puanlamalar yaptıkları bir eğitim verilmiştir. 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. Eğitim sonrasında puanlayıcılar iki gün ile 14 gün arasında değişen sürelerde 104 öğrenciye ait kâğıtları puanlamışlardır.

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. Bu eğitim kapsamında; SOLO taksonomisi ile SOLO taksonomisine dayalı rubrik içeriklerine yer verilmiş ve 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. Araştırma verilerinin toplanması sürecinde takip edilen işlemler Şekil 1'de ayrıca özetlenmiştir.



Şekil 1. Veri Toplama Sürecinde Takip Edilen İşlemler

Veri Analizi

Yukarıda da belirtildiği gibi araştırmanın verileri 104 öğrencinin açık uçlu sekiz matematik sorusuna verdiği cevabın 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, FACETS (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.

Tek Boyutluluk Varsayımı

Araştırma verilerinin tek boyutluluk varsayımını sağlayıp sağlamadığı Açımlayıcı Faktör Analizi (AFA) ile sınanmıştır. Faktör analizi; puanlayıcıların her bir maddeye verdikleri puanların ortalamaları üzerinden gerçekleştirilmiştir. Standart rubrik kullanılarak yapılan puanlamalar için toplam varyansın %31.82'sini açıklayan tek faktörlü bir yapı elde edilmiş ve testte yer alan maddelerin faktör yüklerinin .40 ile .74 arasında değiştiği belirlenmiştir. SOLO taksonomisine dayalı rubrik kullanılarak yapılan puanlamalar için ise testin toplam varyansın %30.84'ünü açıklayan tek faktörlü bir yapıya sahip olduğu ve maddelerin faktör yüklerinin .35 ile .70 arasında sıralandığı saptanmıştır. AFA sonuçlarına göre, tek boyutluluk varsayımının karşılandığı söylenebilir.

Yerel Bağımsızlık

Yerel bağımsızlık tek boyutluluk ile paralel çalışan bir varsayımdır. Dolayısıyla, tek boyutluluk varsayımının karşılandığı durumlarda yerel bağımsızlık varsayımının 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.

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 açık uçlu sekiz soruya verdikleri cevaplar 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 alındığında 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. Bir başka deyişle, 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. Daha sonra, analiz çıktıları puanlayıcı katılığı ve cömertliğine ilişkin alanyazında yer alan ölçütler doğrultusunda incelenmiştir. Puanlayıcı katılığı ve cömertliğini, belirlemeye yönelik olarak incelenen grup düzeyindeki ve bireysel düzeydeki istatistiksel göstergeler (Myford ve Wolfe, 2004) Tablo 7'de sunulmuştur.

Tablo 7. Puanlayıcı Katılığı ve Cömertliğinin İstatistiksel Göstergeleri

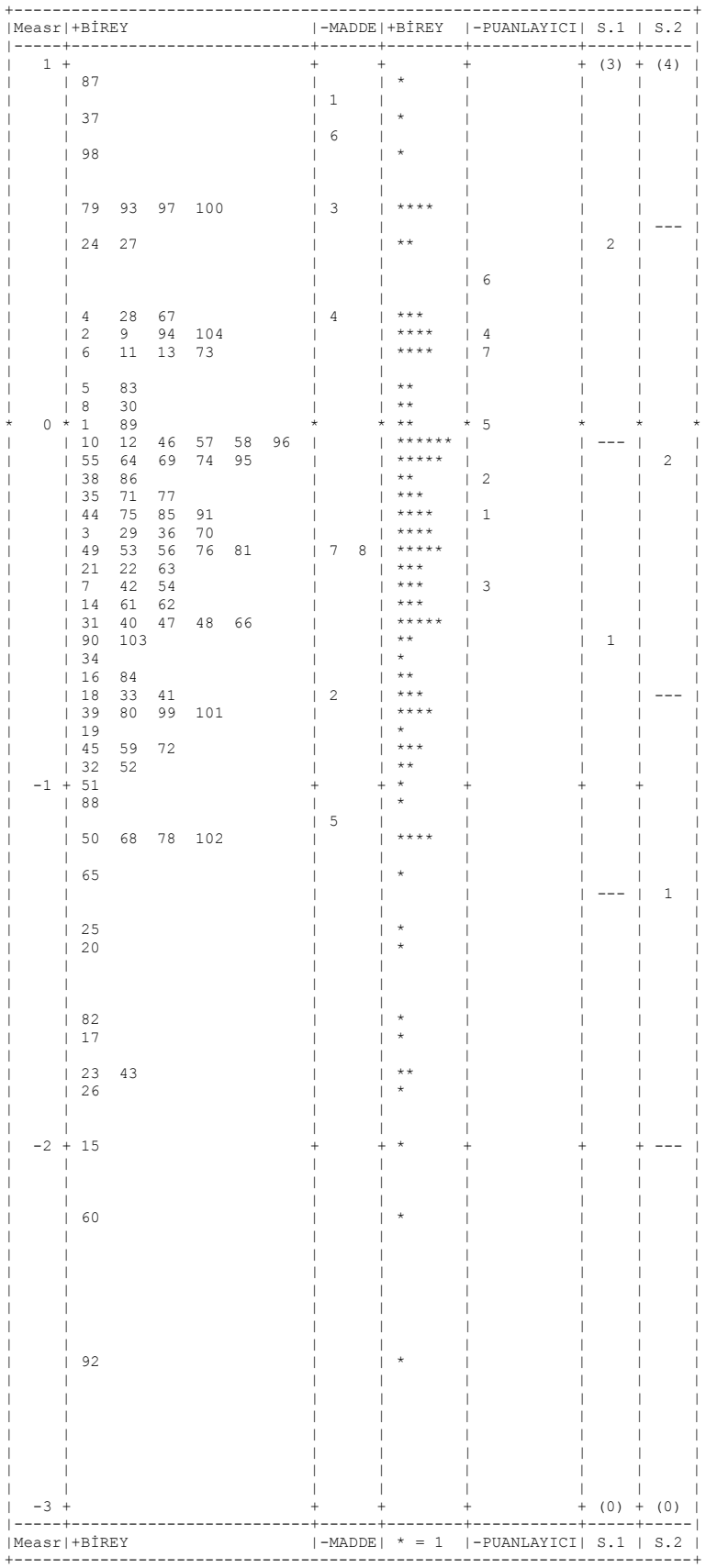
Grup Düzeyinde	**Bireysel Düzeyde**
- Puanlayıcı yüzeyi için hesaplanan Ki Kare değerinin istatistiksel açıdan anlamlı olması	- Puanlayıcılardan herhangi birinin logit cetvel üzerinde diğer puanlayıcılara göre daha farklı bir konumda bulunması
- Puanlayıcı yüzeyine ilişkin ayırma oranı ve güvenirlik indeksinin yüksek olması	- Herhangi bir puanlayıcı için hesaplanan *t*-değerinin istatistiksel açıdan anlamlı olması

Tablo 7'de görüldüğü üzere, puanlayıcı yüzeyine ait Ki Kare testinin anlamlı çıkması puanlayıcı katılığının ve cömertliğinin grup düzeyindeki göstergelerinden ilkidir. Anlamlı olan Ki Kare değeri, puanlayıcılardan en az birinin diğerlerine göre daha katkı ya da daha cömert puanlamalar yaptığını göstermektedir. Puanlayıcı katılığı ve cömertliğinin grup düzeyindeki diğer göstergeleri ayırma oranı ile güvenirlik indeksidir. Güvenirlik indeksi 0 ile 1 arasında değişen değerler alabilirken; ayırma oranı 1 ile sonsuz aralığında uzanmaktadır. Bu iki istatistik farklı metriklerde rapor edilmesine rağmen her ikisi de aynı bilgilerden hesaplanmakta ve belirli bir yüzey için benzer sonuçlara yol açmaktadır. Madde ve birey yüzeyleri söz konusu olduğunda güvenirlik indeksi Cronbach Alpha iç tutarlılık katsayına benzer şekilde yorumlanmaktadır (Bond ve Fox, 2007). Dolayısıyla tıpkı Cronbach Alpha iç tutarlılık katsayısı gibi, madde ve birey yüzeylerine ait güvenirlik indeksi için .70 değerinin ölçüt olarak alınması önerilmektedir (Walker, Engelhard ve Thompson, 2012). Bu ölçütün üzerindeki değerler, yetenek düzeyi farklı olan öğrencilerin başarılı bir biçimde ayırt edilebildiğine ve ölçme aracındaki maddelerin birbirinden bağımsız olarak puanlanabildiğine işaret eder. Puanlayıcı yüzeyi için hesaplanan ayırma oranı ve güvenirlik indeksine ilişkin yüksek değerler ise puanlayıcılar arası uyumun düşük ve farklılaşmanın fazla olduğunu ifade etmektedir. Bu sebeple, puanlayıcı yüzeyine ilişkin güvenirlik indeksi ve ayırma oranının düşük olması istenmektedir. Ancak puanlayıcıların benzer katılık ve cömertlikte puanlama yaptıklarının söylenebilmesi için ayırma oranı ve güvenirlik indeksinin hangi değerin altında olması gerektiğine dair alanyazında net bir ölçüt bulunmamaktadır.

Grup düzeyindeki istatistiksel göstergeler puanlayıcıların katılık ve cömertlikleri arasında fark bulunup bulunmadığını gösterebilse de; eğer bir fark varsa bunun hangi puanlayıcı/puanlayıcılardan kaynakladığı hakkında bilgi vermemektedir. Farka kaynaklık eden puanlayıcının belirlenebilmesi için puanlayıcı katılığı ve cömertliğinin bireysel düzeydeki istatistiksel göstergelerinin incelenmesi gerekir. Puanlayıcılara ait logit ölçüleri, bu logit ölçülerinin ortalaması ve standart hatası kullanılarak hesaplanan *t*-değeri puanlayıcı katılığı ve cömertliğinin bireysel düzeydeki göstergelerinden biridir. Tablo 7'de sunulduğu gibi, puanlayıcıların logit cetvel üzerindeki konumları, puanlayıcı katılığı ve cömertliğinin bireysel düzeydeki istatistiksel göstergelerinden bir diğeridir. Çok yüzeyli Rasch analizinde, tüm değişkenlik kaynakları eşit aralıklı logit ölçeğine dönüştürülmekte ve logit cetvel üzerinde bir arada sunulmaktadır. Puanlayıcıların bu cetvel üzerinde birbirlerine yakın noktalarda kümelenmesi benzer katılılık ve cömertlikte puanlama yaptıklarını gösterir. Logit cetvelin farklı noktalarında bulunmaları ise katılılık ve cömertlikleri yönüyle puanlayıcılar arasında farklılık olduğunu yansıtır.

Bulgular

Bu bölümde, araştırmadan elde edilen bulgulara yer verilmiştir. Çok yüzeyli Rasch analizi çıktıları; birey, madde ve puanlayıcı yüzeylerine ilişkin çok sayıda tablo ile şekil içermektedir. Bununla birlikte, bulguların sunumunda yalnızca puanlayıcı katılığı ve cömertliğinin istatistiksel göstergeleri olabilecek tablo ve şekillere yer vermiştir. İlk olarak, standart rubrikler kullanılarak yapılan puanlamalara ilişkin analiz çıktıları sunulmuştur. Standart rubriklere göre yapılan puanlamaların analiz edilmesiyle elde edilen logit cetvel Şekil 2'de gösterilmiştir.



Şekil 2. Standart Rubriklere Göre Yapılan Puanlamaların Çok Yüzeyli Rasch Modeline Göre Analiz Edilmesiyle Elde Edilen Logit Cetvel

Şekil 2'nin 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 6 numaralı puanlayıcı (.39 logit), en cömert puanlamaların ise 3 numaralı puanlayıcı (-.45 logit) tarafından yapıldığı ortaya çıkmaktadır. Puanlayıcı yüzeyinde logit ölçeğinin negatif ve pozitif ucu boyunca uzanan ölçümlerin elde edilmesi; puanlayıcılar arasında katılık/cömertlikleri yönüyle fark bulunduğuna işaret etmektedir. Ancak bu konuda daha kesin bir kanıya varabilmek için puanlayıcı yüzeyine ait ölçüm raporları incelenmelidir. Puanlayıcı yüzeyine ilişkin ölçüm raporları Tablo 8'de sunulmuştur.

Tablo 8. Standart Rubriklere Göre Yapılan Puanlamaların Çok Yüzeyli Rasch Modeline Göre Analiz Edilmesiyle Puanlayıcı Yüzeyi İçin Elde Edilen Ölçüm Raporları

Puanlayıcı	**Logit Ölçüsü**	**Standart Hata**	**Uyum-İçi**	**Uyum-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			

Tablo 8'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 8'e göre, puanlayıcılara ilişkin logit ölçüleri .39 ile -.45 arasında değişmekte olup puanlayıcıların katılık ve cömertliklerine ilişkin aralık .84 logittir [.39-(-.45)]. Her bir puanlayıcı için rapor edilen logit ölçüsü ile puanlayıcıların logit ölçütlerine ilişkin aralık logit cetvelde de görülmektedir. Tablo 8'de yer alan istatistiklerden biri de, uyum içi ve uyum dışı istatistikleridir. Uyum 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 uyum istatistiklerine ilişkin kabul edilebilir aralığın ne olduğu sorusunun cevaplanması gerekmektedir. Wright ve Linacre (1994), .6 ile 1.4 arasında kalan uyum 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 uyum istatistiklerini kabul edilebilir olarak nitelendirmiştir. Myford ve Wolfe'a (2003) göre, 1.5 ile 2 arasındaki değerler verilerin ölçüm için olumsuz bir etkisi bulunmadığını yansıtırken; 2'nin üstündeki uyum istatistikleri verilerin ölçüm sonuçlarını olumsuz

etkilediğini göstermektedir (Sudweeks, Reeve ve Bradshaw, 2004). Puanlayıcılar için rapor edilen uyum içi ve uyum 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, uyum istatistiklerinin puanlayıcıların hiçbirinde kabul edilebilir aralığın dışında kalmadığı saptanmıştır. Bu bulgu, model ile veri uyumunu olumsuz etkileyen puanlayıcı bulunmadığını yansıtmaktadır.

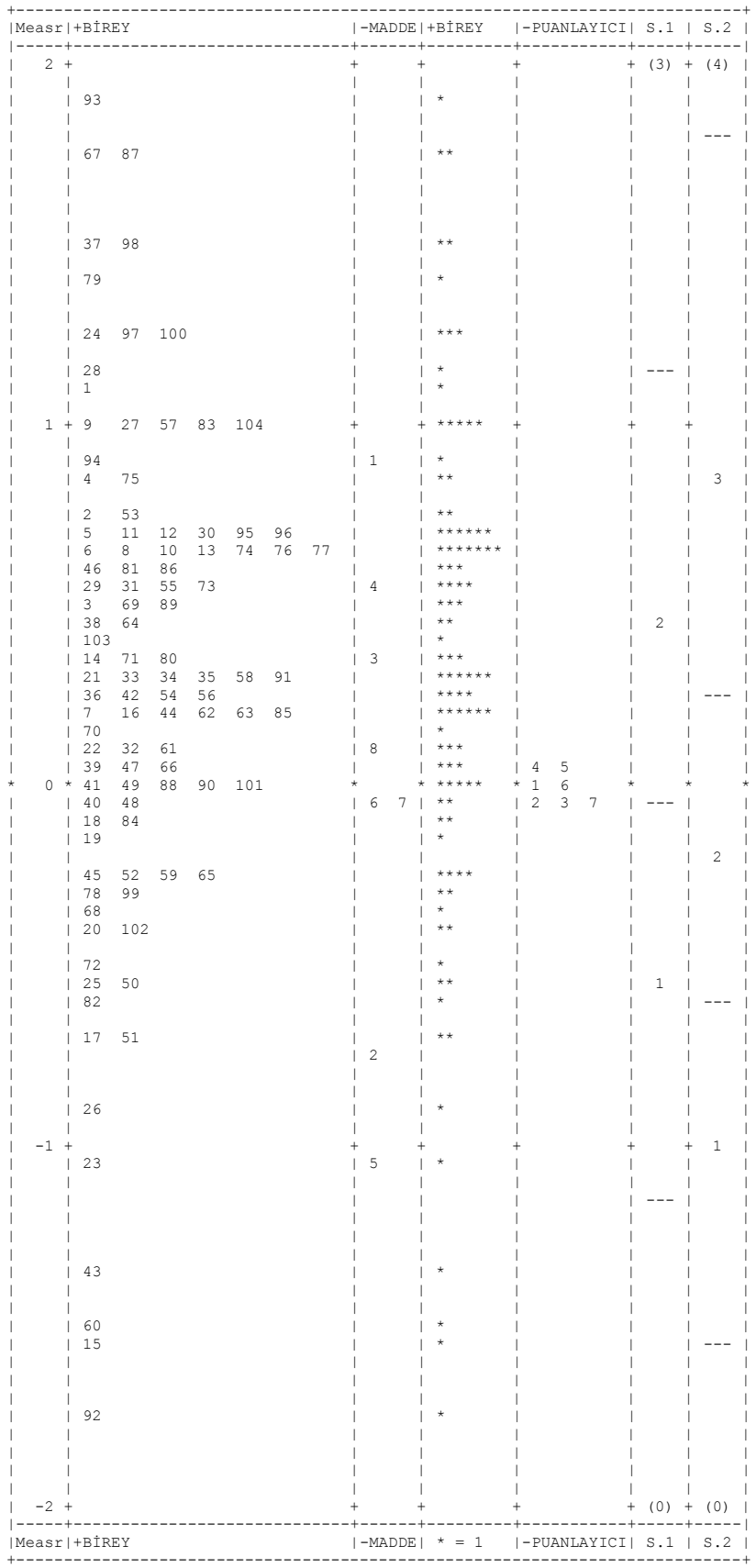
Puanlayıcı yüzeyine ilişkin ayırma oranı ve güvenirlik indeksine bakıldığında evren ve örneklem şeklinde iki farklı model bulunduğu görülmektedir. Linacre'ye (2014) göre, herhangi bir yüzeyin olası bütün bileşenleri model içerisinde yer alıyorsa "*model, evren*" satırındaki ayırma oranı ve güvenirlik 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. Ayırma oranı ve güvenirlik indeksi, farkın güvenirliğine ilişkin istatistiklerdir. Birey yüzeyi için hesaplanan yüksek güvenirlik indeksi yetenek düzeyleri farklı olan öğrencilerin etkili bir biçimde ayırt edilebildiğini yansıtır. Madde yüzeyine ilişkin güvenirlik indeksinin yüksek olması ölçülen özelliğin kavramsal olarak farklı yönlerinin puanlayıcılar tarafından ayırt edilebildiği ortaya koyar. Puanlayıcı yüzeyine ait güvenirlik indeksinin yüksek olması ise puanlayıcıların katılık/cömertlikleri yönüyle farklılık gösterdiği ve puanlayıcılar arası uyumun düşük olduğu anlamına gelir. Çünkü puanlayıcı yüzeyi için hesaplanan güvenirlik indeksi, puanlayıcılar arasındaki güvenilir benzerliği değil; güvenilir farkı göstermektedir (Haiyang, 2010). Buna bağlı olarak; madde ve birey yüzeylerinin aksine, puanlayıcı yüzeyinde ayırma oranı ve güvenirlik indeksinin düşük olması istenmektedir. 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 dağılma 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.

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 logit cetveldir. Puanlayıcılar arasındaki farkın kaynağını, logit cetvelin puanlayıcı sütununda diğer puanlayıcılara göre daha farklı bir noktada bulunan puanlayıcı/puanlayıcıların oluşturduğu kabul edilir (Myford ve Wolfe, 2004). Şekil 2'deki logit cetvele bakıldığında tüm puanlayıcıların logit ölçeğinin farklı noktalarında bulunduğu görülmektedir. 5 numaralı puanlayıcı logit ölçeğinin 0 noktasında yer alırken; 6, 4 ve 7 numaralı puanlayıcılar logit ölçeğinin pozitif ucunda, 2, 1 ve 3 numaralı puanlayıcılar logit ölçeğinin negatif ucunda yer almıştır. Puanlayıcıların logit cetvel üzerinde bulunduğu noktalar; 6, 4 ve 7 numaralı puanlayıcıların puanlamada daha katı; 2, 1 ve 3 numaralı puanlayıcıların ise puanlamada daha cömert davrandığını düşündürmektedir. 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. Daha sonra hesaplanan *t* değeri ile ilgili serbestlik derecesindeki kritik *t* değeri karşılaştırılarak anlamlılık sınaması yapılmaktadır. Araştırmaya yedi puanlayıcı dâhil edildiğinden, serbestlik derecesi 7-1=6 olarak hesaplanmış ve 6 serbestlik derecesi ile .01 düzeyindeki kritik *t* değeri 3.71 olarak belirlenmiştir. Araştırmaya dâhil edilen her bir puanlayıcı için hesaplanan *t* değerleri ve bu değerlerin anlamlılığına ilişkin sonuçlar Tablo 9'da sunulmuştur. Tablo 9'a göre; 6, 4 ve 7 numaralı puanlayıcıların puanlamada daha katı; 2, 1 ve 3 numaralı puanlayıcıların ise daha cömert davrandığı tespit edilmiştir.

Tablo 9. Standart Rubrik Kullanılarak Yapılan Puanlamalarda Katılık ve Cömertlikleri Açısından Puanlayıcılar Arasında Gözlenen Farkın Anlamlılığına İlişkin *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 logit cetvelin 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 logit cetvelin negatif ucunda yer aldığından, bu puanlayıcıların diğer puanlayıcılara göre anlamlı derecede daha cömert puanlamalar yaptığı belirlenmiştir.
P1	-6.00	
P3	-11.25	

Açık uçlu matematik sorularının standart rubriklere göre puanlanması sonucu elde edilen verilerin analizinin ardından, SOLO taksonomisine dayalı rubrikler kullanılarak yapılan puanlamalar analiz edilmiştir. Analiz sonucunda elde edilen logit cetvel Şekil 3'te gösterilmiştir.



Şekil 3. SOLO Taksonomisine Dayalı Rubriklere Göre Yapılan Puanlamaların Çok Yüzeyli Rasch Modeline Göre Analiz Edilmesiyle Elde Edilen Logit Cetvel

Şekil 3'e göre, puanlayıcıların genel olarak logit ölçeğinin sıfır noktasında bulunduğu veya bu noktaya oldukça yakın olduğu saptanmıştır. Puanlayıcıların logit ölçeğinin orta noktasında kümelenmesi, katılık ve cömertlikleri açısından puanlayıcılar arasında önemli bir farkın bulunmadığını düşündürmektedir. Ancak puanlayıcılar arasında anlamlı bir fark olup olmadığını ortaya koyabilmek için puanlayıcı yüzeyine ilişkin ölçüm raporlarının incelenmesi gerekmektedir. Puanlayıcı yüzeyine ilişkin ölçüm raporları Tablo 10'da sunulmuştur.

Tablo 10. SOLO Taksonomisine Dayalı Rubriklere Göre Yapılan Puanlamaların Çok Yüzeyli Rasch Modeline Göre Analiz Edilmesiyle Puanlayıcı Yüzeyi İçin Elde Edilen Ölçüm Raporları

Puanlayıcı	**Logit Ölçüsü**	**Standart Hata**	**Uyum-İçi**	**Uyum-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			

Tablo 10'a 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 uyum içi ve uyum 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, uyum içi ve uyum 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 10'a bakıldığında, puanlayıcı yüzeyine ilişkin ayırma oranının .69 ve güvenirlik indeksinin .32 olarak belirlendiği görülmektedir. 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. Grup düzeyindeki incelemelerin ardından bireysel düzeydeki istatistiksel göstergeler incelenmiştir. Puanlayıcıların logit cetvel üzerindeki konumları, puanlayıcı katılığı ve cömertliğinin bireysel düzeydeki istatistiksel göstergelerinden biridir. Şekil 3'teki logit cetvele 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 11'de sunulmuştur.

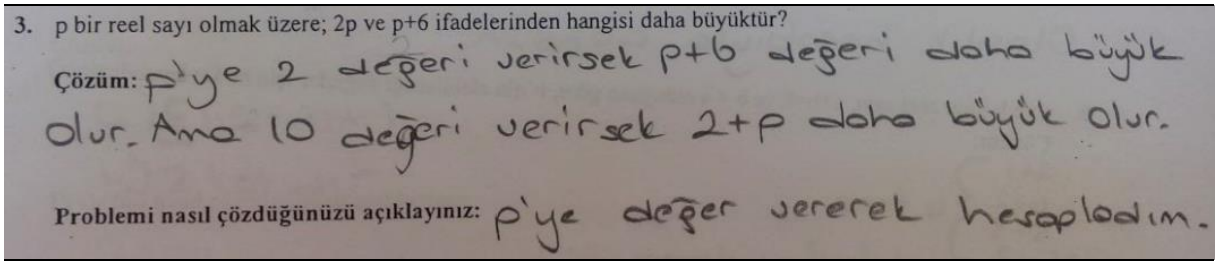
Tablo 11. SOLO Taksonomisine Dayalı Rubrikler Kullanılarak Yapılan Puanlamalarda Katılık ve Cömertlikleri Açısından Puanlayıcılar Arasında Gözlenen Farkın Anlamlılığına İlişkin *t*-testi Sonuçları

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

Tablo 11'e göre, hesaplanan *t* değerleri -1.50 ile 1.50 arasında değişmektedir. Araştırmaya yedi puanlayıcı dâhil edildiğinden, serbestlik derecesi 7-1=6 olarak hesaplanmış ve 6 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.

Standart rubrik kullanarak yapılan puanlamalarda puanlayıcılar arasında anlamlı fark bulunduğu, SOLO taksonomisine dayalı rubrik kullanıldığında ise puanlayıcıların benzer katılık ve cömertlikte puanlamalar yaptığı Tablo 12'de sunulan örnekte de görülebilmektedir. Tablo 12'de matematik başarı testindeki üç numaralı soruya (bk. Ek 1) verilen öğrenci cevaplarından biri, çalışmadaki yedi puanlayıcının bu cevaba atadığı puanlarla birlikte sunulmuştur.

Tablo 12. Puanlayıcıların Standart ve SOLO Taksonomine Dayalı Rubriğe Göre Yaptıkları Örnek Bir Puanlama



	P1	P2	P3	P4	P5	P6	P7
Standart Rubrik	2	0	1	1	1	0	1
SOLO Taksonomisine Dayalı Rubrik	2	2	2	2	2	2	2

Tablo 12 incelendiğinde standart rubrik kullanarak yaptıkları puanlamalarda, puanlayıcıların aynı cevaba oldukça farklı puanlar verdiği anlaşılmaktadır. Aynı cevabı; bir numaralı puanlayıcı standart rubriğin *iyi (2)* kategorisine, üç, dört, beş ve yedi numaralı puanlayıcılar *başlangıç düzeyinde (1)* kategorisine ve altı numaralı puanlayıcı *yetersiz (0)* kategorisine atamıştır. SOLO taksonomisine dayalı rubriklere göre yapılan puanlamalarda ise puanlayıcıların öğrencinin verdiği cevabın SOLO taksonomisinin *çok yönlü yapı* düzeyine karşılık geldiğiyle ilgili hem fikir olduğu ve tüm puanlayıcıların aynı cevaba aynı puanı verdiği görülmektedir.

Tartışma

Bu araştırmada, standart ve SOLO taksonomisine dayalı rubrikler ile puanlanan açık uçlu matematik sorularında puanlayıcı katılığı ve cömertliğinin çok yüzeyli Rasch modeliyle incelenmesi amaçlanmıştır. Araştırmadan elde edilen bulgular; standart rubriklere göre yapılan puanlamalarda puanlayıcılar arası güvenirliğin düşük olduğunu, katılık ve cömertlikleri yönüyle puanlayıcılar arasında anlamlı fark bulunduğunu göstermiştir. Performansa dayalı değerlendirmelerde rubrik kullanımının temel amaçlarından biri, puanlama işleminin kim tarafından yapıldığından bağımsız olarak gerçekleştirilmesini sağlamaktır (Moskal ve Leydens, 2000; Purpura, 2004). Bir başka deyişle, açık uçlu soruların puanlanmasında kullanılan rubriklerin puanlayıcı etkilerini minimum düzeyde tutarak puanlayıcılar arası uyumu arttırması beklenmektedir (Dunbar, Brooks ve Miller, 2006). Ancak araştırmadan elde edilen bulgular standart rubriklerin bu beklentiye yeterince cevap veremediğini yansıtmaktadır. Bu bulgu, Güler ve Gelbal (2010) tarafından yapılan araştırmanın sonuçlarıyla desteklenmektedir. Güler ve Gelbal'ın (2010) yaptığı çalışmada, öğrencilerin açık uçlu matematik sorularına verdikleri cevaplar; herhangi bir taksonomi temele alınmadan hazırlanan holistik rubrikler kullanılarak dört farklı puanlayıcı tarafından puanlanmıştır. Araştırma sonucunda; puanlayıcı güvenirliğinin düşük olduğu, puanlayıcıların katılık ve cömertlikleri arasında anlamlı fark bulunduğu belirlenmiştir. Buna göre, bu çalışmadan elde edilen bulguların Güler ve Gelbal (2010) tarafından yapılan çalışmanın sonuçları ile benzerlik gösterdiği söylenebilir. Ancak bu benzerlik yorumlanırken iki çalışma arasındaki farklılıkların göz ardı edilmemesi gerekir. İlk olarak, Güler ve Gelbal'ın (2010) çalışmasında *genel bir rubrik* geliştirilmiş ve tüm maddelerin puanlanmasında bu genel rubrik kullanılmıştır. Bu çalışmada ise, matematik başarı testindeki her bir soru için ayrı bir rubrik geliştirilmiş ve testteki soruların puanlanmasında bu *göreve özel rubriklerden* yararlanılmıştır. Yine Güler ve Gelbal tarafından yapılan çalışmada, altılı derecelemeye sahip rubrikler kullanılırken, bu araştırmada dörtlü ve beşli derecelemeye sahip standart rubriklerden yararlanılmıştır. Dolayısıyla, ister genel isterse göreve özel olarak geliştirilmiş olsun, ister altılı isterse dörtlü veya beşli dereceleme kullanılmış olsun, standart rubriklerin puanlayıcıların katılık ve cömertlikleri arasındaki farklılıkları giderme konusunda tam anlamıyla etkili olmadığı ifade edilebilir. Bununla birlikte, hem bu araştırmada hem de Güler ve Gelbal'ın (2010) yaptığı çalışmada kullanılan standart rubriklerin holistik yapıya sahip olduğu dikkate alındığında, standart rubriklerin puanlayıcı katılığı ve cömertliği üzerindeki etkisine ilişkin varılan çıkarımın analitik rubrikler için geçerli olmayabileceği gözden kaçırılmamalıdır.

Puanlayıcı katılığı ve cömertliğinin bireysel düzeydeki istatistiksel göstergeleri, araştırmaya dâhil edilen yedi puanlayıcıdan üçü için puanlayıcı katılığının ve üçü için de puanlayıcı cömertliğinin söz konusu olduğunu ortaya koymuştur. Standart rubriklere göre yapılan puanlamalarda, puanlayıcı katılığı ve cömertliğinin araştırmaya katılan puanlayıcıların neredeyse tamamında gözlenen bir puanlayıcı etkisi olması, alanyazındaki kuramsal bilgiler ile örtüşmektedir. Nitekim Cronbach (1990) puanlayıcı katılığı ve cömertliğini, puanlama sürecine karışan en önemli puanlayıcı etkisi olarak ifade etmiştir.

Öğrencilerin açık uçlu matematik sorularına verdikleri cevapların SOLO taksonomisine dayalı rubrikler kullanılarak puanlanması sonucu elde edilen veriler yine çok yüzeyli Rasch modeline göre analiz edilmiştir. Analiz sonuçlarına göre; puanlayıcılar arası uyumun yüksek olduğu, puanlayıcıların katılık ve cömertlikleri arasında anlamlı fark bulunmadığı belirlenmiştir. Bu sonuca dayanarak, SOLO taksonomisine dayalı rubriklerin puanlayıcılar arasındaki farklılıkların giderilmesine yardımcı olduğu ve puanlama işleminin daha objektif bir biçimde gerçekleşmesine katkı sağladığı söylenebilir. SOLO taksonomisinin değerlendirmede kullanılacak ölçütleri açık bir hale getirdiği (Hattie ve Purdie, 1998) ve kolaylıkla anlaşılabilecek düzeylerden oluştuğu (Biggs ve Collis, 1982) şeklindeki kuramsal bilgiler araştırmadan elde edilen bulguları destekler niteliktedir.

Alanyazındaki ampirik çalışmalara bakıldığında SOLO taksonomisine dayalı rubriklerin puanlayıcılar arası güvenirliği arttırdığı yönündeki araştırma bulgusunu destekleyen çalışmalar olduğu gibi; bu bulgu ile çelişen çalışmaların da bulunduğu görülmektedir. Örneğin; Burnett (1999) ve Hundzynski (2008) tarafından yapılan araştırmalarda, SOLO taksonomisi kullanılarak gerçekleştirilen değerlendirmelerde puanlayıcılar arası güvenirlik incelenmiştir. Bu araştırmalarda puanlayıcılar arası güvenirlik katsayıları, sırasıyla .85 ve .87 olarak bulunmuştur. Dolayısıyla, Burnett (1999) 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. Dolayısıyla, sıralanan çalışmalar ile bu araştırmada ulaşılan bulguların aynı eksende olduğu söylenebilir. Ancak, bunu tam bir örtüşme olarak nitelendirmek doğru olmayacaktır. Çünkü puanlayıcılar arasında farklılık olup olmadığı bu araştırmada çok yüzeyli Rasch modeli ile test edilirken; konu ile ilgili önceki çalışmalarda korelasyon analizi ile incelenmiştir. Korelasyon analizi puanlayıcıların değerlendirdikleri bireyler için yaptıkları sıralamanın uyumlu olup olmadığını (göreli uyum) ortaya koymakta, fakat puanlayıcılar arasındaki mutlak uyum hakkında bir bilgi vermemektedir (Goodwin, 2001). Çok yüzeyli Rasch modelinde ise puanlayıcıların katılık ve cömertlikleri arasında fark olup olmadığı değerlendirdikleri bireyler için yaptıkları sıralamalar üzerinden değil; bu bireylere verdikleri puanların gerçek değerleri üzerinden (mutlak uyum) hesaplanmaktadır (Sudweeks vd., 2004). Buna göre; konu ile ilgili önceki çalışmalarda ulaşılan sonuçlar ile bu araştırmadan elde edilen bulgular bir arada ele alındığında SOLO taksonomisine dayalı rubriklerin puanlayıcılar arasındaki hem göreli hem de mutlak uyumu arttırdığı şeklinde bir çıkarıma varılabilir.

Leung (2000) tarafından yapılan çalışmada ise bu araştırmanın bulgularından farklılık gösteren sonuçlara ulaşılmıştır. Leung (2000), SOLO taksonomisine dayalı rubrik kullanılarak yapılan puanlamalarda, puanlayıcı güvenirliğini incelemiş ve puanlayıcılar arası korelasyon katsayısını .49 olarak hesaplamıştır. Leung'a (2000) göre, SOLO taksonomisine dayalı rubrik kullanılarak yapılan değerlendirmelerde puanlayıcılar arası güvenirliğin 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. Bu eğitimler kapsamında yaptırılan örnek puanlamalar, puanlayıcıların rubrik kategorilerine daha aşina hale gelmesini sağlamış olabilir.

Öneriler

Araştırmada ulaşılan sonuçlardan hareketle, hem sınıf içi değerlendirmelerde hem de geniş ölçekli testlerde açık uçlu matematik sorularının puanlanmasında SOLO taksonomisine dayalı rubriklerden yararlanılması önerilebilir. Açık uçlu matematik sorularının puanlanmasında standart rubrikler yerine; SOLO taksonomisine dayalı rubriklerin tercih edilmesi, değerlendirme sonuçlarına karışabilecek puanlayıcı kaynaklı varyansın minimum düzeyde tutulmasına yardımcı olacaktır. Ayrıca SOLO taksonomisi içerikten bağımsız bir model olduğundan (Kanuka, 2011), farklı disiplinlerdeki açık uçlu soruların puanlanmasında da SOLO taksonomisinden yaranılabileceği düşünülmektedir. Bu sonuçlar, çalışmanın giriş kısmında araştırma bulgularının uygulamaya dönük önemli katkılarının olacağına şeklinde ifade edilen beklentiyi doğrulamaktadır.

Araştırma, uygulamaya yönelik önerilerin yanı sıra ileri araştırmalara yönelik bir takım önerileri de beraberinde getirmektedir. Alanyazında, puanlayıcılar arası farklılıkları azaltmada herhangi bir taksonomi temele alınmadan hazırlanan standart rubriklerin mi; yoksa SOLO, Bloom, Fink, Detmer ve Haladyana gibi taksonomilerden herhangi biri temele olarak hazırlanan rubriklerin mi daha etkili olduğu sorusuna cevap olabilecek bir çalışmaya rastlanmamıştır. Standart ve SOLO taksonomisine dayalı rubrikler ile puanlanan açık uçlu matematik sorularında puanlayıcı katılığı ve cömertliğinin karşılaştırmalı olarak incelendiği bu çalışmanın alanyazındaki söz konusu boşluğun giderilmesine katkı sağlayacak bir araştırma olduğu düşünülmektedir. Bununla birlikte; bu araştırma Bloom, Fink, Detmer ve Haladyana gibi diğer taksonomilere dayalı rubriklerin puanlayıcı etkilerini azaltmada standart rubriklere göre daha etkili olup olmadığı sorusunu cevaplamada yetersiz kalmaktadır. Bu kapsamda ileri araştırmalarda, bu taksonomilerden herhangi birine göre hazırlanan rubrikler ile standart rubrikler kullanılarak puanlanan açık uçlu matematik sorularının puanlayıcı etkileri açısından karşılaştırılması önerilebilir. İkinci olarak, bu araştırmada SOLO taksonomisine dayalı rubrikler hazırlanırken Biggs ve Collis (1982) tarafından ileri sürülen beş düzeyli orijinal yapıya sadık kalınmıştır. Ancak alanyazında, SOLO taksonomisinin yedi, sekiz veya dokuz düzeyli olarak yeniden yapılandırıldığı çalışmalara da (Burnett, 1999; Chan vd., 2002) rastlanmaktadır. Alanyazındaki 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 katılığı ve cömertliğinin incelenmesi yerinde olacaktır. Üçüncü olarak, bu araştırmada incelenen puanlayıcı etkileri puanlayıcı katılığı ve cömertliği ile sınırlı tutulmuştur. Daha sonra yapılacak çalışmalarda; merkeze yönelme etkisi, halo etkisi, yanlılık ve tutarsızlık gibi diğer puanlayıcı etkileri de incelenebilir. Çalışmanın, 104 öğrencinin açık uçlu sekiz matematik sorusuna verdikleri cevabın 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 bir veri kaynağı üzerinde yapılması önerilebilir. Son olarak, bu araştırmada, standart ve SOLO taksonomisine dayalı rubriklerin puanlayıcı etkilerini azaltma konusundaki işlevselliği açık uçlu matematik soruları üzerinden incelenmiştir. Benzer çalışmaların farklı dersler için de yapılması, araştırmadan elde edilen bulguların genellenebilirliğine katkı sunması açısından oldukça önemlidir.

Kaynakça

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

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

Barkaoui, K. (2008). *Effects of scoring method and rater experience on ESL essay rating processes and outcomes* (Yayımlanmamış doktora tezi). University of Toronto, Canada. http://search.proquest.com/docview/304360302/fulltextPDF/4AEA7C68D8F945FEPQ/1?accountid=15780 adresinden erişildi.

Biggs, J. B. ve Collis, K. F. (1982). *Evaluating the quality of learning: The SOLO taxonomy*. Academic 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. http://mimoza.marmara.edu.tr/~hakkoc/yayin2008_bingolbali_ozmantar_akkoc_usos.pdf adresinden erişildi.

Bond, T. G. ve Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Erlbaum.

Brabrand, C. ve Dahl, B. (2009). Using the SOLO taxonomy to analyze competence progression of university science curricula. *Higher Education*, *58*(4), 531-549. doi:10.1007/s10734-009-9210-4

Brentari, E. ve Golia, S. (2008). Measuring job satisfaction in the social services sector with the Rasch model. *Journal of Applied Measurement*, *9*(1), 45-56. http://www.unibs.it/sites/default/files/ricerca/allegati/10061.pdf adresinden erişildi.

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. doi:10.1080/03069889908256291

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

Chan, C. C., Tsui, M. S., Mandy, Y. C. ve 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. doi:10.1080/0260293022000020282

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

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

Çetin, B., Boran, A. ve Yazıcı, N. (2014). Fizik eğitiminde başarının ölçülmesinde SOLO taksonomisine göre hazırlanan rubriklerin incelenmesi. *Bayburt Üniversitesi Eğitim Fakültesi Dergisi*, *9*(2), 32-71. http://edergi.bayburt.edu.tr/index.php/befd/article/view/9/6 adresinden erişildi.

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

Dunbar, N. E., Brooks, C. F. ve 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. doi:10.1007/s10755-006-9012-x

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. doi:10.1207/s15434311laq0203_2

Erden, M. ve Akman, Y. (2011). *Eğitim psikolojisi*. Ankara: Arkadaş Yayınevi.

Farrokhi, F., Esfandiari, R. ve 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. http://www.idosi.org/wasj/wasj15(IPLL)11/12.pdf adresinden erişildi.

Goodwin, L. D. (2001). Interrater agreement and reliability. *Measurement in Physical Education and Exercise Science*, *5*(1), 13-34. doi:10.1207/S15327841MPEE0501_2

Gronlund, N. E. (1998). *Assessment of student achievement*. Boston: Allyn and Bacon.

Güler, N. (2008). *Klasik test kuramı, genellenebilirlik kuramı ve Rasch modeli üzerine bir araştırma* (Yayımlanmamış doktora tezi). Hacettepe Üniversitesi, Sosyal Bilimler Enstitüsü, Ankara, Türkiye.

Güler, N. ve Gelbal, S. (2010). Klasik test kuramı ve çok değişkenlik kaynaklı Rasch modeli üzerine bir çalışma. *Eğitim Araştırmaları Dergisi*, *38*, 108-125. http://www.aniyayincilik.com.tr/main/pdfler/38/7_guler_nese.pdf adresinden erişildi.

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. http://www.celea.org.cn/teic/90/10060807.pdf adresinden erişildi.

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

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

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.

Hundzynski, C. (2008). *Elementary teachers in a science inquiry study group: Concerns, uses, and reflections* (Yayımlanmamış doktora tezi). Fordham University, New York, ABD. http://search.proquest.com/docview/304641444/previewPDF/AB40E91C649C453CPQ/1?accountid=15780 adresinden erişildi.

İlhan, M. (2016). Açık uçlu sorularla yapılan ölçmelerde klasik test kuramı ve çok yüzeyli Rasch modeline göre hesaplanan yetenek kestirimlerinin karşılaştırılması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 31*(2), 346-368. doi:10.16986/HUJE.2016015182

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

Jurdak, M. (1991). Van Hiele levels and the SOLO taxonomy. *International Journal of Mathematical Education in Science and Technology*, *22*(1), 57-60. doi:10.1080/0020739910220109

Kan, A. (2005). Yazılı yoklamaların puanlanmasında puanlama cetveli ve yanıt anahtarı kullanımının (farklı) puanlayıcı güvenirliğine etkisi. *Eğitim Araştırmaları Dergisi*, *19*, 207-219. http://www.ejer.com.tr/0DOWNLOAD/pdfler/tr/821760610.pdf adresinden erişildi.

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. https://www.edam.com.tr/kuyeb/pdf/en/567aeeee08a7e62db0b82fd5312c9d7baneng.pdf adresinden erişildi.

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. doi:10.1007/s12528-011-9049-4

Kind, P. M. (1999). Performance assessment in science-What are we measuring?. *Studies in Educational Evaluation*, *25*(3), 179-194. doi:10.1016/S0191-491X(99)00021-8

Koretz, D., McCaffrey, D., Klein, S., Bell, R. ve Stecher, B. (1992). *The reliability of scores from the 1992 Vermont portfolio assessment program* (Center fort he Study the Evaluation Tech Rep No: 350). Santa Monica, CA: Rand Institute on Education and Training.

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

Lake, D. (1999). Helping Students to go SOLO: Teaching critical numeracy in the biological sciences. *Journal of Biological Education*, *33*(4), 191-198. doi:10.1080/00219266.1999.9655664

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. doi:10.1023/A:1008937007674

Lian, L. H. ve Idris, N. (2006). Assessing algebraic solving ability of form four students. *International Electronic Journal of Mathematics Education*, *1*(1), 55-76. http://www.mathedujournal.com/dosyalar/a4.pdf adresinden erişildi.

Lian, L. H. ve Yew, W. T. (2012). Assessing algebraic solving ability: A theoretical framework. *International Education Studies*, *5*(6), 177-188. doi:10.5539/ies.v5n6p177

Linacre, J. M. (2014). *A user's guide to FACETS Rasch-model computer programs.* http://www.winsteps.com/a/facets-manual.pdf adresinden erişildi.

Lucas, U. ve 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. doi:10.1007/s10734-009-9218-9

Matematik Öğretmenleri Ulusal Konseyi. (2000). *Principles and Standards for School Mathematics*. Reston, VA: Author. http://www.nctm.org/store/Products/Principles-and-Standards-for-School-Mathematics-(Book-and-E-Standards-CD)/ adresinden erişildi.

McBee, M. M. ve 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. doi:10.1207/s15324818ame1102_4

McNamara, T. F. (1996). *Measuring second language performance*. London and New York: Longman.

Milli Eğitim Bakanlığı. (2007). *Matematik öğretmen kılavuz kitabı*. Ankara: Devlet Kitapları Müdürlüğü.

Milli Eğitim Bakanlığı. (2009). *İlköğretim matematik dersi 6-8. sınıflar öğretim programı*. http://ttkb.meb.gov.tr/program2.aspx adresinden erişildi.

Milli Eğitim Bakanlığı. (2013). *Temel eğitimden ortaöğretime geçişle ilgili sıkça sorulan sorular*. http://www.meb.gov.tr/duyurular/duyurular2013/bigb/tegitimdenoogretimegecis/MEB_SSS_20_09_2013.pdf adresinden erişildi.

Mohd Nor, N. ve Idris, N. (2010). Assessing students' informal inferential reasoning using SOLO taxonomy based framework. *Procedia Social and Behavioral Sciences*, *2*(2), 4805-4809. doi:10.1016/j.sbspro.2010.03.774

Mooney E. S. (2002). A framework for characterizing middle school students' statistical thinking, *Mathematical Thinking and Learning*, *4*(1), 23-63. doi:10.1207/S15327833MTL0401_2

Moore, B. B. (2009). *Consideration of rater effects and rater design via signal detection theory* (Yayımlanmamış doktora tezi). Columbia University, New York. http://search.proquest.com/docview/304862541 adresinden erişildi.

Moskal, B. M. ve Leydens, J. A. (2000). Scoring rubric development: validity and reliability. *Practical Assessment, Research & Evaluation*, *7*(10), 71-81. http://pareonline.net/getvn.asp?v=7&n=10 adresinden erişildi.

Mulqueen, C., Baker D. ve Dismukes, R. K. (2000, Nisan). *Using multifacet Rasch analysis to examine the effectiveness of rater training*. 15th Annual Conference for the Society for Industrial and Organizational Psychology (SIOP) konferansında sunulmuş bildiri, New Orleans. http://www.air.org/files/multifacet_Rasch.pdf adresinden erişildi.

Myford, C. M. ve 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. ve Wolfe, E. W. (2004). Detecting and Measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applıed Measurement*, *5*(2), 189-227. http://jimelwood.net/students/grips/tables_figures/myford_wolfe_2004.pdf adresinden erişildi.

National Assessment Governing Board. (2002). *Mathematics framework for the 2003 national assessment of educational progress.* http://academic.wsc.edu/faculty/jebauer1/mat645/framework_03.pdf adresinden erişildi.

Öğrenci Seçme ve Yerleştirme Merkezi. (2013). *Açık uçlu sorularla deneme sınavının uygulanması*. http://www.osym.gov.tr/belge/1-19410/acik-uclu-sorularla-deneme-sinavinin-uygulanmasi-311201-.html adresinden erişildi.

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. http://mimoza.marmara.edu.tr/~hakkoc/yayin2008_ozmantar_bingolbali_akkoc_usos.pdf adresinden erişildi.

Palm, T. (2008). Performance assessment and authentic assessment: A conceptual analysis of the literature. *Practical Assessment, Research & Evaluation*, *13*(4), 1-11. http://pareonline.net/getvn.asp?v=13&n=4 adresinden erişildi.

Purpura, J. E. (2004). *Assessing grammar*. Cambridge University Press.

Rembach L. ve Dison, L. (2016). Transforming taxonomies into rubrics: Using SOLO in social science and inclusive education. *Perspectives in Education*, *34*(1), 68-83. http://scholar.ufs.ac.za:8080/xmlui/bitstream/handle/11660/3838/persed_v34_n1_a6.pdf?sequence=1&isAllowed=y adresinden erişildi.

Romagnano, L. (2001). The myth of objectivity in mathematics assessment. *Mathematics Teacher*, *94*(1), 31-37. http://www.peterliljedahl.com/wp-content/uploads/Myth-of-Objectivity.pdf adresinden erişildi.

Romberg, T. E. ve Wilson, L. D. (1992). Issues related to development of authentic assessment system for school mathematics. T. A. Romberg (Ed.). *Reform in school mathematics and authentic assessment* içinde (s. 1-18). Albany: State University of New York Press.

Saal, F. E., Downey, R. G. ve Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, *88*(2), 413-428. doi:10.1037/0033-2909.88.2.413

Stecher, B. (2010). *Performance assessment in an era of standards-based educational accountability.* Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education. https://scale.stanford.edu/system/files/performance-assessment-era-standards-based-educational-accountability.pdf adresinden erişildi.

Sudweeks, R. R., Reeve, S. ve 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. doi:10.1016/j.asw.2004.11.001

Tan, Ş. (2015). *Öğretimde ölçme ve değerlendirme KPSS el kitabı*. Ankara: Pegem Akademi Yayıncılık.

Tekin, H. (2009). *Eğitimde ölçme ve değerlendirme*. Ankara: Yargı Yayınevi.

Toffoli, S. F. L., Andrade, D. F. ve Bornia, A. C. (2016). Evaluation of open items using the many-facet Rasch model. *Journal of Applied Statistics*, *43*(2), 299-316, doi:10.1080/02664763.2015.1049938

Walker, E. R., Engelhard, G. ve Thompson, N. J. (2012). Using Rasch measurement theory to assess three depression scales among adults with epilepsy. *Seizure*, *21*(6), 437-443. doi:10.1016/j.seizure.2012.04.009

Wiseman, C. S. (2012). Rater effects: Ego engagement in rater decision-making. *Assessing Writing, 17*(3), 150-173. doi:10.1016/j.asw.2011.12.001

Woodward, J., Monroe, K. ve Baxter, J. (2001). Enhancing student achievement on performance assessments in mathematics. *Learning Disability Quarterly, 24*(1), 33-46. doi:10.2307/1511294

Wright, B. D. ve Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement: Transactions of the Rasch Measurement SIG*, *8*(3), 370.

Yazıcı, N. (2013). *Başarının ölçülmesinde SOLO taksonomiye dayalı hazırlanan rubrik kullanımının etkisinin karşılaştırmalı olarak incelenmesi* (Yayımlanmamış yüksek lisans tezi). Kahramanmaraş Sütçü İmam Üniversitesi, Sosyal Bilimler Enstitüsü, Kahramanmaraş, Türkiye.

Zhu, X. (2009). *Assessing fit of item response models for performance assessments using bayesian analysis* (Yayımlanmamış doktora tezi). University of Pittsburgh, Pittsburgh, ABD. http://d-scholarship.pitt.edu/10162/1/XiaowenZhu_ETD2009_Final.pdf adresinden erişildi.

Ek 1

Açık Uçlu Matematik Sorusu

p bir reel sayı olmak üzere; $2p$ ve $p+6$ ifadelerinden hangisi daha büyüktür?

Standart Rubrik

Puanlama Ölçütleri	
3 puan *Çok İyi*	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 *İyi*	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 *Başlangıç Düzeyinde*	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 getirememiştir. -Problemi çözmek için uygun strateji ile başlangıç yapmanın dışında problemin çözümüne yönelik doğru işlemler yapılamamıştır. Yapılan işlemlerde önemli hatalar bulunmaktadır.
0 puan *Yetersiz*	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.

SOLO Taksonomisine Dayalı Rubrik

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.