Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü
Eğitim Bilimleri Anabilim Dalı
Eğitimde Ölçme ve Değerlendirme Bilim Dalı

KLASİK TEST KURAMI GENELLENEBİLİRLİK KURAMI VE RASCH MODELİ ÜZERİNE BİR ARAŞTIRMA

Neşe Güler

Doktora Tezi

Ankara, 2008

KLASİK TEST KURAMI GENELLENEBİLİRLİK KURAMI VE RASCH MODELİ ÜZERİNE BİR ARAŞTIRMA

Neşe Güler

Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü
Eğitim Bilimleri Anabilim Dalı
Eğitimde Ölçme ve Değerlendirme Bilim Dalı

Doktora Tezi

Ankara, 2008

KABUL VE ONAY

Neşe Güler tarafından hazırlanan "Klasik Test Kuramı Genellenebilirlik Kuramı ve Rasch Modeli Üzerine Bir Araştırma" başlıklı bu çalışma, 10.06.2008 tarihinde yapılan savunma sınavı sonucunda başarılı bulunarak jürimiz tarafından Doktora Tezi olarak kabul edilmiştir.

Prof. Dr. Yaşar Baykul (Başkan)

Doç. Dr. Selahattin Gelbal (Danışman)

Prof. Dr. Nizamettin Koç

Prof. Dr. Aysun Umay

Doç. Dr. Şener Büyüköztürk

Yukarıdaki imzaların adı geçen öğretim üyelerine ait olduğunu onaylarım.

Prof. Dr. İrfan Çakın
Enstitü Müdürü

BİLDİRİM

Hazırladığım tezin tamamen kendi çalışmam olduğunu ve her alıntıya kaynak gösterdiğimi taahhüt eder, tezimin/raporumun kağıt ve elektronik kopyalarının Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü arşivlerinde aşağıda belirttiğim koşullarda saklanmasına izin verdiğimi onaylarım:

- [x] Tezimin/Raporumun tamamı her yerden erişime açılabilir.
- [ ] Tezim/Raporum sadece Hacettepe Üniversitesi yerleşkelerinden erişime açılabilir.
- [ ] Tezimin/Raporumun .....yıl süreyle erişime açılmasını istemiyorum. Bu sürenin sonunda uzatma için başvuruda bulunmadığım takdirde, tezimin/raporumun tamamı her yerden erişime açılabilir.

10.06.2008

Neşe Güler

CANIM BABACIĞIMA VE EŞİME

TEŞEKKÜR

Çalışmalarım boyunca değerli yardım ve katkılarıyla beni yönlendiren, desteklerini esirgemeyen, yapıcı önerileri ve yorumlarıyla bu çalışmanın tamamlanmasında büyük katkıları olan danışmanım ve hocam Doç. Dr. Selahattin GELBAL'a teşekkürlerimi sunarım.

Tez konumu belirlemeye çalıştığım ilk aşamadan itibaren her türlü heyecanımı paylaşan, yardımları ve önerileriyle çalışmama katkıda bulunan hocam Doç. Dr. Hülya KELECİOĞLU'na, desteğini her zaman hissettiğim, her soru ve sıkıntımda ilgisini ve yardımlarını eksik etmeyen hocam Dr. Nuri DOĞAN'a teşekkür ederim.

Görüşleriyle ve eleştirileriyle çalışmama değerli katkılarda bulunan jüri üyelerim Prof. Dr. Yaşar BAYKUL, Prof. Dr. Nizamettin KOÇ, Prof. Dr. Aysun UMAY ve Doç. Dr. Şener BÜYÜKÖZTÜRK'e şükranlarımı sunarım.

Tez yazım sürecinde çeşitli şekillerde desteğini gördüğüm arkadaşım Arş. Gör. Sevda ÇETİN'e ve sabırla bana katlanan ve hep yanımda olan oda arkadaşlarım Dr. Necla EKİNCİ ve Arş. Gör. Özlenen ÖZDİYAR'a teşekkür ederim.

Yüreğimde ve yanımda sevgisiyle ve desteğiyle, çalışmamın ortaya çıkmasındaki sabır dolu yardımlarıyla hiçbir zaman beni yalnız bırakmayan canım eşime çok teşekkür ederim.

Adını sayamadığım, sevginin ve bilginin paylaştıkça arttığına inanmış, sevgisini ve bilgisini benimle paylaşan herkese teşekkürlerimle...

ÖZET

GÜLER, Neşe. *Klasik Test Kuramı Genellenebilirlik Kuramı ve Rasch Modeli Üzerine Bir Araştırma*, Doktora Tezi, Ankara, 2008.

Bu araştırmada, 2007 yılında yapılan matematik başarısının ölçülmesiyle elde edilen puanlara klasik test kuramı, genellenebilirlik kuramı ve çok değişkenlik kaynaklı Rasch ölçme modeli uygulanmıştır. Uygulanan bu üç kurama göre puanların güvenirlikleri hesaplanmış ve üç kuramdan elde edilen sonuçlar karşılaştırılmıştır.

Araştırmanının ilk aşamasında, TIMSS-1999'da yer alan açık uçlu matematik sorularından 24'ü, 2007 yılı bahar döneminde 203 öğrenciye uygulanmıştır. Daha sonra, bu öğrencilerin verdikleri cevaplar dört puanlayıcı tarafından holistik rubrik ile puanlanmıştır. Araştırmanının ikinci aşamasında, elde edilen puanların güvenirliği farklı kuramlara göre incelenmiştir. Klasik test kuramında Cronbach alfa güvenirlik katsayısı, puanlayıcılar arası uyumun belirlenmesinde Kedall'ın konkordans katsayısı, puanlayıcılar arası korelasyon katsayısı ve puanlayıcıların verdikleri puanların ortalamaları arası fark olup olmadığı F testi ile araştırılmıştır. Genellenebilirlik kuramında, b x g x p tümüyle çaprazlanmış desen kullanılarak genellenebilirlik ve güvenirlik katsayıları hesaplanmıştır. Çok değişkenlik kaynaklı Rasch ölçme modeli ile birey, puanlayıcı ve madde boyutlarına ilişkin ayrı ayrı güvenirlik hesaplamaları yapılmıştır.

Bu araştırma neticesinde, klasik test kuramına göre matematik başarısının ölçülmesiyle elde edilen puanların iç tutarlılığının 0.92 gibi oldukça yüksek bir değer olduğu görülmüştür. Puanlayıcılar arası uyumun belirlenmesinde Kendall'ın konkordans katsayısı 0.52 olmakla birlikte puanlayıcılar arası korelasyon katsayıları 0.90 ile 0.97 arasında değişen değerler göstererek puanlayıcıların verdikleri puanlar arasında anlamlı bir ilişki olduğu sonucuna varılmıştır. Ancak F testi ile elde edilen sonuçlara göre puanların ortalamaları arasında farklılık olduğu belirlenmiştir. Genellenebilirlik kuramına göre matematik başarısının ölçülmesiyle elde edilen puanların genellenebilirlik katsayısı 0.92 ve güvenirlik katsayısı 0.90 bulunmuştur. Puanlayıcı

değişkenlik kaynağının toplam varyansı açıklama yüzdesi 2.1 ile oldukça düşük bir değer göstermiştir. Çok değişkenlik kaynaklı Rasch ölçme modeline göre öğrenci boyutunun güvenirliği 0.95 olarak hesaplanmıştır. Bu modele göre puanlayıcılar arası güvenirlik ise 0.99 olarak bulunmuştur.

Elde edilen tüm bu sonuçlara göre, 2007 yılında uygulanan matematik başarısını belirlemek için kullanılan ölçme aracının, öğrencilerin matematik başarısını belirlemede güvenilir sonuçlar verdiği görülmüştür. Matematik başarısının belirlenmesinde yer alan dört puanlayıcının puanları ortalamaları arasında fark olmakla birlikte, birbirleriyle uyumlu puanlama yaptıkları belirlenmiştir.

Araştırma ile matematik başarısının ölçülmesinde güvenirliğin belirlenmesinde yararlanılacak kuramların hangisinin seçileceği, elde edilen puanların hangi amaç için kullanılacağına bağlı olarak değişebileceği görülmüş, ancak araştırma sonuçlarına göre, matematik başarısının ölçülmesinde güvenirliğin belirlenmesinde en az iki kuramdan yararlanmanın daha uygun olacağı sonucuna varılmıştır.

Anahtar Sözcükler
Klasik test kuramı, genellenebilirlik kuramı, çok değişkenlik kaynaklı Rasch ölçme modeli

ABSTRACT

GÜLER, Neşe. A Research On *Classical Test Theory Generalizability Theory and Rasch Model*, Ph. D. Dissertation, Ankara, 2008.

In this study, classical test theory, generalizability theory and multi facet Rasch measurement model were applied to the scores which were obtained from mathematics performance measurement in 2007. According to these three theories, inter-rater reliability was figured out and the results were compared each other.

In first step of this study, 24 open-ended questions of 1999-TIMSS were applied to 203 students in 2007 spring semester. Later, the students' responds were scored by four raters. In second step of this study, the reliability of the scores was analyzed in the view of different theory. In the classical test theory, Cronbach alpha reliability coefficient, Kendall's concordance coefficient for inter-rater reliability and correlation coefficients of four raters' scores were calculated and it was investigated whether there was a difference among the means of raters' scores with F test. In generalizability theory, by using p x t x r (all facet cross with each other) design, generalizability and dependability coefficient were calculated. With multi facet Rasch measurement model, the reliability was figured out for person, task and rater facets separately.

In the results of this study, according to classical test theory inter consistency of the mathematic performance measurement was found as 0.92. Although Kendall's concordance coefficient for four raters was obtained as 0.52, correlation coefficients for four raters were different values between 0.90 and 0.97. Thus, it was concluded that there is a statistically significant correlation between raters. However, according to F test it was found that there was a difference between the means of the raters' scores. According to the generalizability theory, the generalizability and the dependability coefficient of the mathematic performance measurement were 0.92 and 0.90, respectively. Variance due to raters accounts for only %2.1 of the total variance which suggests that very little of the variability found in the model for differences among raters who scored the mathematic performance measurement. According to multi facet

Rasch measurement model, the reliability of person facet was 0.95 and the reliability of rater facet was 0.99.
According to all results, for determining of the students' success in mathematics the reliability of the mathematics performance measurement which was applied in 2007 was found as very high. Although there was a difference between the means of the raters' scores it was obtained that the four raters scored the students consistently.

With this study, it was seen that the theory to be selected for the determination of the reliability of the performance measurement depended upon the purpose for which the scores obtained would be used. However, it was concluded that for determination of the reliability of the performance measurement, at least two theories should rather be used.

Key Words
Classical test theory, Generalizability theory, Multi facets Rasch measurement model

İÇİNDEKİLER

TABLOLAR DİZİNİ

ŞEKİLLER DİZİNİ

BÖLÜM I

GİRİŞ

Bilimsel çalışmaların gerçekleştirilmesinde ölçmenin tartışılmaz bir yeri vardır. Ölçme, kuramlarda yer alan teorik ilişkilerin, gerçek hayatta test edilmesini sağlayarak, bilimsel çalışmaların vazgeçilmez bir parçası olmaktadır. Ölçme ile elde edilen veriler temel alınarak ilke ve genellemelere varılırken, bu genellemelerin farklı koşullarda sınanmasıyla kanunlara ulaşmak yine ölçmeyle mümkün olmaktadır. Buna ek olarak, bilimin işlevsel sürecinde yer alan, aynı amaca yönelik oluşturulan kuramların aynı şartlar altında kıyaslanarak, bilgi sağlama güçlerinin, çalışan ve çalışmayan yanlarının saptanması, daha pratik ve doğru olanının belirlenmesi, kuramların benzer ve farklı yönlerinin açığa çıkarılmasında ölçme önemli bir yere sahiptir.

Pek çok bilim dalında olduğu gibi eğitimin de hem kuramsal hem deneysel bir yönü bulunmaktadır. Eğitimde, zeka, tutum, öğrenme gibi yapıların incelenmesinde, içeriklerinin açıklanmasında ve bireylerin bu niteliklerinin sayılarla ifade edilmesinde ölçmenin ve ölçme araçlarının kullanılması zorunluluğu vardır (Baykul, 2000). Ölçme ve ölçme araçları eğitim sürecinin vazgeçilmez bir parçası olan değerlendirmenin temelini oluşturmaktadırlar.

Eğitimde, eğitim programının işlevini yerine getirebilme derecesinin sınanmasında, eğitim yöntem ve stratejilerinin etkililik düzeylerinin belirlenmesinde, öğrenci başarısının, akademik tutumunun saptanmasında, öğrencilerin öğrenmedeki güçlü ve zayıf yönlerinin tespit edilmesinde ölçme ve değerlendirmeden yararlanılmaktadır. Bir başka deyişle eğitimin girdi, süreç ve ürün boyutlarının tümünde ölçme ve değerlendirmeye ihtiyaç duyulmaktadır. Bu noktada ölçme ile değerlendirme arasındaki farkın anlaşılmasında fayda vardır. Ölçme, bir özelliğin gözlenerek gözlem sonuçlarının sayı ya da sembollerle ifade edilmesi (Baykul, 2000) olarak tanımlanırken, değerlendirme daha geniş kapsamlı bir süreç olup; ölçme sonuçları (ölçüm), ölçüt ve karar verme basamaklarını içermektedir. Ölçme sonuçlarının bir ölçütle karşılaştırılarak karara varılması olarak tanımlanan değerlendirmenin doğru yapılmış olmasında, kullanılan ölçütün uygunluğu yanısıra, ölçme sonuçlarının güvenilir ve geçerli

olmasının çok büyük önemi bulunmaktadır. Değerlendirme sonuçlarının isabetli olma derecesini arttırabilmek için yapılan ölçme işlemlerinde kullanılan ölçme araçlarının güvenirliğinin ve geçerliğinin olabildiğince yüksek olması istenir. Bu kapsamda, eğitimde yer alan çeşitli ölçme araçları arasında en çok kullanılan ikisi çoktan seçmeli testler ve performansın ölçülmesidir (Alharby, 2006).

Çoktan seçmeli testler, bir maddenin cevabını, verilen ikiden fazla seçenek arasından bulmayı gerektiren sorulardan oluşan ölçme araçlarıdır. Çoktan seçmeli testlerde, her bir soru bütününe "madde" adı verilmektedir. Her bir madde, kök ve seçenekler olmak üzere iki bölümden oluşmaktadır. Maddenin kökü, sorunun sorulduğu bölümü, seçenekler ise soruya ilişkin cevabın seçilmesi için sunulan olası durumları içermektedir. Seçenekler, biri doğru cevap olmak üzere, diğerleri "çeldiriciler" olarak isimlendirilen yanlış cevaplardan oluşmaktadırlar. Çeldiriciler; maddede yer alan soruyla ölçülmek istenilen davranışa ya da beceriye sahip olmayan ya da az sahip olan öğrenicileri yanıltmak üzere hazırlanmış ifadelerdir. Çoktan seçmeli testlerde, cevap öğrencinin kendisi tarafından verilmez, cevap zaten test maddesi içinde yer almaktadır. Öğrenci test süresinin çok büyük bir kısmını test maddelerini okumak ve doğru cevabı bulmak için kullandığından, yazmaya ayrılan süre çok az olup, çoktan seçmeli testlerde çok sayıda madde kullanılarak pek çok davranış ölçülebilir. Böylece çoktan seçmeli testlerde yer alan soru sayısının çokluğu, öğrencilerin çok daha geniş bir kapsamda ölçülebilmesine imkan tanır. Ölçme aracında yer alan soru sayısının fazla olması, ölçmenin hem güvenirliğini hem de kapsam geçerliğini arttıran bir unsurdur. Çoktan seçmeli testlerin uygulanması kolay olduğu gibi puanlanması da kolay ve aynı zamanda doğası gereği tamamıyla objektiftir. Bir başka deyişle, çoktan seçmeli testlerde cevaplama işlemi, seçenekler arasından seçimle yapıldığından, cevapların puanlaması puanlayıcıdan puanlayıcıya ya da zamandan zamana farklılık göstermez. Hem soru sayısının fazla olması hem de puanlamasının objektif olması çoktan seçmeli testleri diğer ölçme araçlarından daha güvenilir kılan en önemli özellikler arasında yer alır.

Çoktan seçmeli testlerin tüm bu avantajları yanında bazı sınırlılıkları da bulunmaktadır. Bunlardan biri, tahminle doğru cevaplamanın mümkün olmasıdır. Çoktan seçmeli test maddelerinde cevaplayıcılar, gösterilmesi beklenilen davranış ya da becerinin dışında, doğru cevabı tahminle bulabilirler. Cevaplayıcının maddenin doğru cevabını bilerek ya

da tam anlamıyla emin olarak değil tahminle bulmasına şans başarısı adı verilir (Rogers, 1997). Şans başarısı cevaplayıcının rasgele, ipucuyla ya da bir miktar bilgiye sahip olarak cevap vermesiyle ortaya çıkabilir. Rasgele şans başarısında, cevaplayıcının maddenin doğru cevabına ilişkin hiç bir fikri yoktur, maddeyi tamamıyla rasgele cevaplandırır. İpucuyla ortaya çıkan şans başarısında, cevaplayıcı doğru cevabı bilmediği halde, maddede ya da seçeneklerde istemeden verilmiş olan bir ipucundan yararlanarak maddeyi doğru cevaplandırır. Bunlara ek olarak şans başarısı, cevaplandırıcının maddeye ilişkin bir miktar bilgiye sahip olması ve bu bilgiden yola çıkarak tahminde bulunup maddeyi cevaplandırmasıyla da söz konusu olabilir (Rogers, 1997). Şans başarısıyla ortaya çıkan en önemli psikometrik sorun, test puanlarındaki varyansı arttırmasıdır. Bu durum maddelerin güvenirlik ve geçerliğinin düşmesine sebep olur. Çoktan seçmeli testlerde diğer bir sınırlılık da kopya çekmeye daha elverişli olmasıdır. Cevaplayıcılar için çoktan seçmeli testlerde kopya çekmek, açık-uçlu sorularda kopya çekmekten daha kolay görünmektedir (Alharby, 2006).

Bir başka sınırlılık, çoktan seçmeli testlerin test-tekniği becerisini de barındırıyor olmasıdır. Cevaplayıcılar, çoktan seçmeli maddelerle kendilerinden göstermeleri beklenen gerekli bilgi ya da beceriye sahip olmaksızın, doğru cevabı seçme şanslarını arttırıcı beceriye sahipseler, bu durumda doğru cevabı bulabilmektedirler. Böylece çoktan seçmeli maddelerle ölçülmek istenilen asıl özelliğin dışında farklı bir özelliğin ölçülüyor olması ölçme aracının yapı geçerliğini düşürmektedir.

Yukarıda açıklanan tüm sınırlılıklar ikinci derece sınırlılıklar olarak tanımlanır. Çünkü bu sınırlılıkların üstesinden gelinebilmektedir (Alharby, 2006). Örneğin; şans başarısı, madde tepki kuramının 3-parametreli lojistik modeli (3-PL) kullanılarak kontrol altına alınabilir. 3-parametreli lojistik model, cevaplayıcının maddede göstereceği performansın maddenin üç özelliğinden etkilendiği varsayımına dayanmaktadır: madde güçlük düzeyi (β_i) ve madde ayırıcılık düzeyi (α_i) parametreleri ve çok düşük yetenek düzeyindeki cevaplayıcının maddeyi doğru cevaplama olasılığını gösteren şans başarısı (γ_i) parametresi olmak üzere; 3-parametreli lojistik model aşağıda verilen eşitlik ile ifade edilmektedir:

$$P(X_{is} = 1|\theta_s, \beta_i, \alpha_i, \gamma_i) = \gamma_i + (1 - \gamma_i)\frac{e^{[\alpha_i(\theta_s - \beta_i)]}}{1 + e^{[\alpha_i(\theta_s - \beta_i)]}}$$

Eşitlikte yer alan $X_{is} = 1$ ifadesi, s cevaplayıcısının i maddesine verdiği cevabı (cevap doğru olduğunda bire eşit olacaktır) ve θ_s s cevaplayıcısının yetenek düzeyini göstermektedir (Rogers, 1997).

Ayrıca şans başarısı, "düzeltme formülü" kullanılarak engellenmeye çalışılabilir. Düzeltme formülüne göre; cevaplayıcının doğru cevapladığı madde sayısı "D", yanlış cevapladığı madde sayısı "Y" ve maddenin seçenek sayısı "S" ile gösterilirse, cevaplayıcının şans başarısına göre düzeltilmiş puanı P= D-[Y/(S-1)] formülüyle hesaplanabilir (Rogers, 1997). Düzeltilmiş puanın doğruluğu ancak üç varsayımın gerçekleşmesiyle mümkündür: 1. Öğrenci doğrularını cevap kağıdına işaretlerken kaydırma yapmamıştır. 2. Öğrencinin yaptığı tüm yanlışlar şanssızlığından dolayıdır. 3. Öğrencinin şansla cevapladığı bir maddedeki her bir seçeneği işaretleme olasılığı birbirine eşttir. Bu varsayımlar sağlanmadığında, düzeltme formülü kullanılarak hesaplanan düzeltilmiş puan hatalı olacaktır (Baykul, 2000). Bu formül istenilenden fazla düzeltme yapmaya sebep olabilmektedir. Örneğin bir öğrenci 30 maddelik beş seçenekli bir çoktan seçmeli testte 25 maddeyi yanlış 5 maddeyi doğru cevaplandırmışsa, bu durumda öğrencinin düzeltilmiş puanı, 5-(25/5-1)= -1.25 olacaktır. Ancak, bu öğrencinin hiç bir bilgiye sahip olmadığı düşünüldüğünde bile öğrencinin alabileceği en düşük puan 0'dır. Bu nedenle düzeltme formülünün kullanılması sınıf içi ölçmelerde çok uygun olmamaktadır (Atılgan, Kan ve Doğan, 2006).

Şans başarısını azaltmanın bir başka yolu, sınav yönergesinde, öğrencilerin sadece bildikleri, cevabından emin oldukları soruları cevaplamaları gerektiğini belirtmek olabilir. Şans başarısını azaltmanın diğer bir yolu ise maddedeki seçenek sayısını arttırmaktır. Ancak seçenek sayısını arttırmaya çalışırken uygun olmayan, zorunlulukla oluşturulmuş seçenekler eklemek, işlevini yerine getirmeyen boşuna yazılmış seçenekler olabilir. Şans başarısı yukarıdaki gibi bazı yöntemler ile kontrol altına alınmaya çalışılırken; çoktan seçmeli testlerin bir diğer dezavantajı olan kopya çekmek, paralel test formlarının kullanılması, ortamın testin yapılması için uygun hale getirilmesi gibi

stratejiler kullanılarak engellenebilir ya da azaltılabilir. Çoktan seçmeli testlerin bu ikinci derece sınırlılıkları dışında çok daha ciddi başka sınırlılıkları bulunmaktadır (Alharby, 2006). Pek çok önemli eğitimsel performansı, yeteneği, ya da beceriyi çoktan seçmeli testlerle yeterli derecede ölçmek mümkün olmamaktadır. Çoktan seçmeli testlerle ileri düzeydeki bilişsel davranışların ölçülmesi oldukça zordur. Öğrenme teorisinde önemli bir yere sahip olan bilişsel düzeyler Bloom (1956) tarafından basitten karmaşığa doğru aşamalı olarak altı sınıfta toplanmıştır. Bu sınıflamanın ilk basamağı *bilgi*; kavram, olgu ya da herhangi bir nesnenin görünce hatırlanması, sorulunca derste öğretildiği ya da kitaptaki şekliyle aynen söylenmesi olarak tanımlanır. İkinci basamakta yer alan *kavrama*; bilgi basamağında kazanılan davranışların öğrenci tarafından özümsenmesini, bilgiyi anlamını bozmadan kendi ifadeleriyle dile getirlebilmesini içerir. Bu basamakta öğrenci, bilgi farklı biçimlerde verildiğinde de bilgiyi hatırlayabilmekte, bu bilgiye dayalı çıkarımlarda bulunabilmektedir. Bilişsel davranış düzeylerinin üçüncü aşamasında *uygulama* yer alır. Bu aşama, bilgi ve kavrama düzeyinde kazanılan davranışların yeni bir problem durumunun çözümünde işe koşulmasını gerektirir. Bir sonraki aşama olan *analiz*de ise, bilgiyi, kendisini oluşturan parçalara ayrıştırmak, bütün ile parçaları arasındaki ilişkiyi kullanabilmek ön plana çıkmaktadır. Beşinci basamakta yer alan *sentez*; bilgi parçalarını tamamıyla yaratıcı bir düşünceyle bir araya getirerek yeni ve özgün bir ürün ortaya çıkarmayı kapsar. Bilişsel davranışların son düzeyi *değerlendirme*dir. Değerlendirme basamağında, belirli standartlar veya ölçütler geliştirerek, fikirler, materyaller, yöntemler, teknikler vb. niteliksel ve niceliksel özellikleri açısından karşılaştırılarak bir yargıya varılır, değerlendirilir (Atılgan, Kan ve Doğan, 2006; Sönmez, 2003).

Yapılan araştırmalar, çoktan seçmeli testlerle yukarıda sıralanan Bloom'un taksonomisinde yer alan alt seviyedeki becerilerin ölçülebildiğini göstermektedir. Ancak sentez ve değerlendirme düzeyindeki davranışların çoktan seçmeli test maddeleriyle ölçülmesi oldukça zordur (Alharby, 2006). Özellikle müzik, resim yeteneği, makale yazımı, iletişim ve ifade becerisi gibi bazı akademik alanlardaki performansların ölçülmesinde çoktan seçmeli maddelerin kullanılması yeterli ve uygun olmamaktadır. Diğer bazı alanlarda, örneğin; fen bilimleriyle ilgili derslerdeki deneylerde öğrencilerin deney düzeneğini kurabilme, deneyi yapabilme ve sonuçlarını yorumlayabilme gibi sürece odaklı deneysel becerilerinin ölçülmesi, matematik

alanında öğrencilere sunulan yeni problemleri çözme süreçlerinin gözlenmesi, çoktan seçmeli test maddeleriyle yeterince mümkün olmamaktadır.

Çoktan seçmeli testlerle ilgili bir diğer ciddi problem ise öğrenmeye olan etkisidir. Eğitime ilişkin yapılan ölçmelerin en önemli amaçlarından biri öğretim programının gelişimine yardımcı olmak iken, çoktan seçmeli testlerin kullanımı programın tüm diğer amaçlarının üstünde "testlerin beklentilerine" odaklanan öğretim programlarının oluşmasına sebep olabilmektedir. Böylece pek çok ailenin, öğrencinin ve öğretmenlerin asıl ilgilendiği nokta, öğrencilerin testlerle ölçülen becerileri kazanıp kazanmadıklarından çok, bu testlerdeki başarıları ya da başarısızlıkları olmaktadır. Bu durum, daha önemli becerilerin öğretilmesinde başarısızlık yaşanmasına sebep olabilmektedir (Alharby, 2006). Çoktan seçmeli testlerin bu sınırlılıkları, öğrencilerin davranışlarının ölçülmesinde başka yöntemlerin aranmasına neden olmuştur. Bu yöntemlerden birisi performansın ölçülmesidir.

Performansın ölçülmesinde, öğrencilerin akademik bilgi ve becerilerinin gerçek yaşamda karşılaşılan problemlere uygulayabilme düzeylerini belirleyebilmek amaçlanır. Performansın ölçülmesi; öğrencinin resim çizimindeki süreci, bir kimya deneyinde, deney düzeyini hazırlama ve deneyi yapmadaki becerilerini gözlemlemeyi, kompozisyon yazma becerisinin ölçülmesini ya da bir matematik problemini çözerken izlediği yolları adım adım takip etmeyi mümkün kılar. Tüm bu durumların ortak özelliği verilen görevlerin gerçek yaşamdaki durumlarla ilgili olmasıdır. Baker, O'Neil ve Linn (1993) performansın ölçülmesinin altı genel özelliğini sıralamıştır. Performansın ölçülmesi,

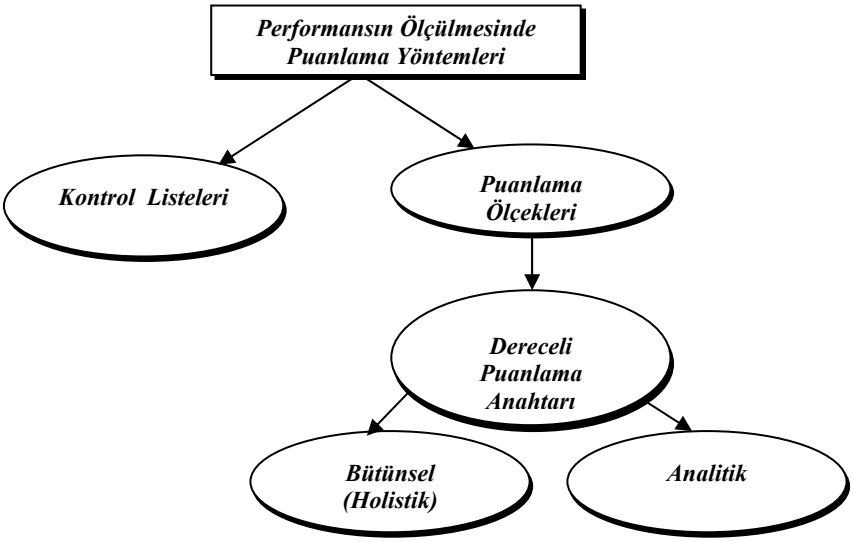
1. Çoktan seçmeli maddelerden oluşmaz, açık-uçlu sorular biçimindedir.
2. Daha yüksek bilişsel beceri düzeylerine odaklanma eğilimindedir.
3. Öğrencilerin maddeleri ya da görevleri doğru cevaplandırmaları için gerekli stratejileri doğrudan uygulamalarını gerektirir.
4. Çoğunlukla tamamlamak için daha fazla zamanın gerektiği karmaşık görevlerden oluşmaktadır.
5. Bireysel olarak ya da geniş gruplara uygulanabilir.

6. Genellikle verilen görevlerin cevaplandırılmasında öğrencilere daha fazla özgürlük tanımaktadır.

Böylece performansın ölçülmesi, öğrencilerin öğrendiklerine ilişkin çok daha geniş açıdan bilgi sahibi olabilmeye, öğrencilerin zayıf ve güçlü yönlerinin belirlenmesine olanak sağlar. Performansın ölçülmesinin tüm bu özellikleri kapsamında yer alan avantajlarından biri, çoktan seçmeli testlerle ölçmenin zor olduğu yüksek düzeydeki bilişsel beceriler için kolaylıkla kullanılabilmesidir. Bir diğer avantajı da, çoktan seçmeli testlere göre ölçülmek istenilen özellikler daha açık olarak ölçülebildiği için, performansın ölçülmesinin çoktan seçmeli testlerden daha geçerli ölçmeler vermesidir (Alharby, 2006).

1.1. PERFORMANSIN ÖLÇÜLMESİNDE PUANLAMA YÖNTEMLERİ

Performansın ölçülmesiyle ilgili alanyazında yer alan önemli noktalardan biri puanlama yöntemleridir. Puanlayıcıların puanlamada kullandıkları yöntem, ölçmenin geçerliğini olduğu kadar güvenirliğini de etkilemektedir. Performansın ölçülmesinde en çok kullanılan yöntemler, başlıca kontrol listeleri ve puanlama ölçekleri olarak sınıflandırılabilir. Puanlama ölçeklerinin en çok bilineni ve kullanılanılanı dereceli puanlama anatarları (rubrikler) da, bütünsel (holistik) ve analitik olarak ikiye ayrılabilir. Performansın ölçülmesinde kullanılan bu yöntemler Şekil 1'de gösterildiği gibi resmedilebilir (Mertler, 2001).



Şekil 1. Performansın Ölçülmesinde Kullanılan Puanlama Yöntemleri

1.1.1. Dereceli Puanlama Anahtarı (Rubrikler)

Dereceli puanlama anahtarları, öğrencilerin gösterdikleri performansların sürecini ve sonucunu analiz edebilmek için öğretmenler ya da ölçme uzmanları tarafından geliştirilen açıklayıcı puanlama şemalarıdır (Brookhart, 1999). Puanlama öncesi bu tür bir şemanın hazırlanmış olması, puanlamanın subjektif olduğu performansın ölçülmesinde yer alan görev ya da maddelerin daha objektif puanlanabilmesini sağlar. Dereceli puanlama anahtarlarında her bir puan kategorisine ilişkin cevap özelliklerinin tanımlanmış olması, birbirinden bağımsız puanlayıcıların bir soruya verilen cevaba aynı puanı verme olasılığını arttırır (Moskal, 2000). Performansın ölçülmesinde yer alan görev ya da maddelerin dereceli puanlama anahtarları kullanılmadan puanlanmasında, örneğin; bir öğrencinin bir matematik sorusuna verdiği cevap için on puan üzerinden yedi puan alması gibi, öğrencinin performansının hangi aşamasında nasıl bir başarı ya da başarısızlık gösterdiğine ve bir sonraki performansında nasıl bir gelişme göstermesi gerektiğine ilişkin açıklayıcı bir bilgi bulunmaz. Halbuki dereceli puanlama anahtarlarıyla yapılan puanlamalarda, öğrencinin her bir görev ya da maddedeki performansının hangi düzeyde olduğu, performansını geliştirmek için hangi kriterleri sağlaması gerektiği görülebilmektedir.

Bir performansın ölçülmesinde dereceli puanlama anahtarının uygun bir puanlama yöntemi olup olmadığına karar vermek, hangi ders ya da sınıf düzeyi için kullanılacağından çok yapılan ölçmenin amacına bağlıdır (Moskal, 2000). Dereceli puanlama anahtarları, grup aktivitelerinin, projelerin ya da sözel sunumların puanlanmasının yanı sıra İngilizce, matematik ve fen derslerinde, lise ve öncesi sınıf düzeylerinde kullanılabilirler. Öğrencinin gelişmesi için geri bildirimde bulunmak ya da öğrencinin belirli kriterlere göre nereye ulaştığına dair açık bir resminin çizilmesi isteniyorsa dereceli puanlama anahtarlarının kullanılması uygun olacaktır. Dereceli puanlama anahtarları bütünsel (holistik) ve analitik olmak üzere iki sınıfa ayrılmaktadır.

1.1.1.1. Bütünsel (Holistik) Dereceli Puanlama Anahtarları

Bütünsel dereceli puanlama anahtarları ilk kez 1960'larda ortaya çıkmış ve ilk olarak genel izlenimle puanlama yöntemi olarak tanımlanmıştır (Nitko, 2001). Bütünsel dereceli puanlama anahtarları öğrencilerin her bir görev ya da maddede göstermiş

oldukları performansın bir bütün olarak, alt bileşenlerine ayrılmaksızın puanlanmasında kullanılır. Bütünsel dereceli puanlama anahtarlarıyla yapılan puanlamada, her bir görev ya da maddenin bütününde gösterilen performans yüksek ise bazı bileşenlerindeki hatalar görmezden gelinebilir (Mertler, 2001). Bütünsel dereceli puanlama anahtarlarıyla puanlamada, puanlayıcı her bir görev ya da maddeyi bir bütün olarak inceleyerek, bu bütüne ilişkin bir puan vermektedir. Bir başka deyişle, bütünsel dereceli puanlama anahtarlarıyla puanlama tek boyutta yapılmaktadır. Pek çok araştırmacı bu yöntemin, puanlayıcının daha subjektif olmasına sebep olabilecek daha gereksiz ayrıntılara takılmaksızın puanlama yapmasına imkan verdiğini savunmaktadır. Örneğin Klein ve arkadaşları (1998), bu puanlama yaklaşımının bütünün parçalarından daha büyük önem taşıdığı durumlarda yani cevapların organizasyon ve ikna etme gücü gibi genel özelliklerinin kalitesinin değerlendirildiği durumlarda daha uygun olduğunu vurgulamışlardır. Tablo 1'de bütünsel dereceli puanlama anahtarlarına ilişkin bir örnek yer almaktadır (Mertler, 2001).

Tablo 1. Bütünsel Dereceli Puanlama Anahtarları İçin Bir Örnek

Puanlar	Tanımlar
5	Problemin *tümüyle* anlaşılmış olduğu gösterilmiştir. Göreve ya da maddeye ilişkin *tüm gereklilik* cevapta yer almıştır.
4	Problemin *oldukça* anlaşılmış olduğu gösterilmiştir. Göreve ya da maddeye ilişkin *tüm gereklilik* cevapta yer almıştır.
3	Problemin *kısmen* anlaşılmış olduğu gösterilmiştir. Göreve ya da maddeye ilişkin *çoğu gereklilik* cevapta yer almıştır.
2	Problemin *çok az bir kısmının* anlaşılmış olduğu gösterilmiştir. Göreve ya da maddeye ilişkin *çoğu gereklilik* cevapta yer almamıştır.
1	Problemin *anlaşılmamış* olduğu gösterilmiştir.
0	Cevap verilmemiştir /görev hiçbir şekilde gerçekleştirilmemiştir.

1.1.1.2. Analitik Dereceli Puanlama Anahtarları

Analitik dereceli puanlama anahtarlarında, bütünsel olanların aksine, öğrencinin her bir görev ya da maddede gösterdiği performans öncelikle alt bileşenlerine ayrılarak

puanlanır. Görev ya da maddeye ilişkin puan bu bileşenlerden alınan puanların toplamıdır. Bir başka deyişle, analitik dereceli puanlama anahtarlarıyla puanlama çok boyutta yapılmaktadır (Mertler, 2001). Bir görev ya da maddede gerçekleştirilen performansın her bir alt boyutunda öğrencinin güçlü ve zayıf yönleri belirlenebilir. Pek çok araştırmacıya göre, bir alanın alt alanlarına ayrılarak yapılan puanlamada, puanlayıcıların subjektifliğinden daha az etkilenme olmakta ve puanlayıcılar arası değişkenlik daha azalmaktadır (Alharby, 2006). Tablo 2'de analitik dereceli puanlama anahtarlarına ilişkin bir örnek yer almaktadır (Mertler, 2001),

Tablo 2. Analitik Dereceli Puanlama Anahtarları İçin Bir Örnek

	1 Puan	2 Puan	3 Puan	4 Puan	Toplam
Ölçüt 1	Performansın başlangıç seviyesi gösterilmiştir.	Performansta üst düzeye doğru bir gelişme gösterilmiştir.	Üst düzey bir performans gösterilmiştir.	En üst düzeyde performans gösterilmiştir.	Puan
Ölçüt 2	Performansın başlangıç seviyesi gösterilmiştir.	Performansta üst düzeye doğru bir gelişme gösterilmiştir.	Üst düzey bir performans gösterilmiştir.	En üst düzeyde performans gösterilmiştir.	Puan
Ölçüt 3	Performansın başlangıç seviyesi gösterilmiştir.	Performansta üst düzeye doğru bir gelişme gösterilmiştir.	Üst düzey bir performans gösterilmiştir.	En üst düzeyde performans gösterilmiştir.	Puan
Ölçüt 4	Performansın başlangıç seviyesi gösterilmiştir.	Performansta üst düzeye doğru bir gelişme gösterilmiştir.	Üst düzey bir performans gösterilmiştir.	En üst düzeyde performans gösterilmiştir.	Puan

Performansı ölçmek isteyen kişi pek çok faktörü düşünerek bütünsel ya da analitik dereceli puanlama anahtarlarından hangisini kullanmanın daha uygun olacağına karar verir. Clauser (2000)'a göre, göz önünde bulundurulması gereken en önemli faktörlerden biri, ölçmede yer alan görevlerin (soruların) türüdür. Örneğin, fen bilgisi dersinde öğrencinin yapacağı bir deneyde göstereceği performansın, deney düzeneğini oluşturabilme, analizleri doğru yapabilme, doğru yorum ve sonuçlara ulaşabilme gibi alt bileşenlerinin ölçüleceği durumlarda analitik dereceli puanlama anahtarı kullanmak daha uygun olabilir (Linn ve Gronlund, 2000). Diğer taraftan performansın alt bileşenlerinin ölçülmesine gerek olmadığı performansın ölçülmesinde ise holistik dereceli puanlama anahtarı tercih edilebilir. Dereceli puanlama anahtarı seçimindeki bir diğer önemli faktör, ölçmenin yapılacağı kişi sayısıdır. Bu açıdan bakıldığında çok sayıda bireyin katıldığı performans ölçmelerinde, zaman açısından tasarruf sağlayacağı

için bütünsel dereceli puanlama anahtarı kullanmak kolaylık sağlayacaktır. Dereceli puanlama anahtarlarından hangisinin daha uygun olacağının belirlenmesindeki bir başka faktör ise performansın ölçülmesiyle elde edilen sonuçların hangi amaç için kullanılacağıdır. Özellikle sınıf içi ölçmelerde olduğu gibi detaylı geri bildirimlere ihtiyaç duyulan durumlarda analitik dereceli puanlama anahtarı tercih edilebilir. Analitik dereceli puanlama anahtarıyla sınıf çalışmalarının güçlü ve zayıf yönleri kolaylıkla saptanabilmekte ve sonraki öğretim sürecinin belirlenmesinde yol gösterici olabilmektedir. Değer biçmeye dönük bir değerlendirme yapılacaksa, performansa ilişkin bütüncül bir puan elde edilmek isteniyorsa bütünsel dereceli puanlama anahtarı kullanmak uygun olabilir. Puanlayıcıların dereceli puanlama anahtarı kullanmasına yönelik bilgilendirilmeleri ve uzmanlaştırılmaları hangi tür için daha kolay olacaksa o dereceli puanlama anahtarı türü tercih edilebilir. Ancak şunu belirtmekte yarar vardır ki analitik dereceli puanlama anahtarı için eğitim vermek bütünsel dereceli puanlama anahtarından çok daha zaman alıcı ve zahmetli olmaktadır (Boring, 2002). Son olarak, eğer performansın alt bileşenlerinin teoride taşıdıkları önem, farklılık gösteriyorsa analitik dereceli puanlama anahtarı bu farklılıklara göre ağırlıklandırmayla puanlama yapılmasına imkan vermektedir (Boring, 2002). Bu ise bir ya da daha fazla alt bileşenin diğerlerinden daha kritik olduğu durumlarda oldukça önem taşımaktadır.

1.2. PERFORMANSIN ÖLÇÜLMESİNDE GEÇERLİK

Geçerlik en genel tanımıyla, bir ölçme aracının sadece o ölçme aracıyla ölçülmek istenilen özelliği ölçmesi, başka özellikleri karıştırmamasıdır (Baykul, 2000). Son yirmi yılda geçerlik ile ilgili alanyazında yer alan açıklamalar giderek gelişen bir değişim göstermektedir. 1974'te Amerikan Psikolji Derneği'nin yayınladığı *Psikoloji ve Eğitim Testleri için Standartlar*'da (Akt. Fine, 1995) geçerlik; "test puanlarının ya da diğer ölçme araçlarından elde edilen sonuçların yorumlanmasındaki uygunluk" olarak tanımlanmıştır. Geçerliğin tanımlanmasında aşağıda belirtilen dört özelliğin göz önünde bulundurulması gerektiği önerilmektedir (Linn ve Gronlund, 2000):

1) Geçerlik, ölçme sürecinin kendisinin değil, belirli bir grubun ölçülmesi süreci sonunda elde edilen puanların yorumlanmasının uygunluğudur. Geçerlik bazen "ölçme aracının geçerliği" olarak ifade edilse de "ölçme sonuçlarının

kullanımının ya da yorumlanmasının geçerliği" olarak ifade edilmesi daha uygun doğru olacaktır.

2) Geçerlik, ya hep ya hiç şeklinde ifade edilecek bir olgu olmayıp bir derece olarak ifade edilebilir. Kısacası, ölçme sonucu elde edilen puanlar geçerlidir ya da geçerli değildir demek yerine geçerliği yüksektir ya da düşüktür şeklinde ifade etmek çok daha doğru olmaktadır.
3) Geçerlik, ölçme sonuçlarının belirli bir kullanımı ya da yorumlanması için belirlenebilecek bir özelliktir. Hiç bir ölçme sonucu tüm durumlar için geçerli olamaz. Örneğin; dört işlem yapabilme becerisine yönelik hazırlanmış bir ölçme aracı bu amaç için yüksek geçerliğe sahip olabilirken, matematiksel olarak mantık yürütme becerisini ölçme amacına yönelik bir ölçme durumu için çok ölçme amacı için geçerli bir ölçme durumu olmamaktadır.
4) Geçerlik için belirli ve tek bir tanım vermektense; geçerliği ölçme aracının ölçme amacına uygun olduğunu destekleyecek kanıtların elde edilmesine yönelik tanımların verilmesi daha doğru olacaktır.

Yukarıda ifade edilen dördüncü maddeden de anlaşılacağı üzere geçerlik farklı açılardan farklı tanımlamalar gerektirir. Geçerlik birbiriyle ilişkili üç ayrı açıdan kavramsal olarak tanımlanabilir: *Yapı geçerliği, kapsam geçerliği*, *ölçüt (kriter) geçerliği*. *Yapı geçerliği* ölçme aracıyla ölçülmek istenilen hipotetik yapının pratikte ölçülebilme derecesini, *kapsam geçerliği* ölçme aracının ölçülmesi amaçlanan kapsam alanını temsil edebilme derecesini, *ölçüt geçerliği* de ölçme sonuçlarının, ikinci bir ölçüt ile ilişkisinin derecesini ifade etmektedir. Daha sonraki beş yıl içinde yapılan eğitim ve psikolojiye ilişkin ölçme çalışmaları bu görüşlere ortak fikirler üzerinde birleşmiştir (Cronbach, 1980; Linn, 1980). Halen, *Psikoloji ve Eğitim Testleri için Standartlar* 1985 versiyonu (Akt. Fine, 1995), geçerliğe ilişkin bu üç bakış açısını ele almaktadır. Ölçme sonuçları birden fazla şekilde kullanılacak ya da yorumlanacaksa, tüm kullanımlar ya da yapılan yorumların her biri geçerli olmalıdır.

1.3. MATEMATİK ÖĞRETİMİNDE PERFORMANSIN ÖLÇÜLMESİ

Günümüzde matematik eğitimiyle ilgili olarak yer alan yaygın görüş; matematik öğretiminin, birçok beceri ve kavramın özellikle hatırlanmasından çok daha üst düzeyde

yeterlilikleri içermesi gerektiğidir. Bu da öğrencinin kendisinde var olan bilgilerle yeni bilgileri bütünleştirerek yeni anlamlar oluşturmasını mümkün kılacak aktif bir süreci kapsar. Matematiksel fikirlerin anlaşılması sadece hatırlama ile sınırlandırılamaz, çok daha üst düzey bilişsel becerilere sahip olmayı gerektirir. Matematik; öğrenme, keşfetme, kontrol etme, problem çözme, sonuçları açıklama, niteliksel ve niceliksel sebepler sunmayı içerir. Matematiğin ayrıca duyuşsal bir boyutu da bulunmaktadır. Matematiğin duyuşsal boyutu, matematiğin toplumdaki yerini değerlendirme, diğer disiplinlerle bağlantısını idrak edebilme, matematik yeteneğine ilişkin güveni geliştirebilmeyi kapsar.

Tüm bu açılardan, matematiksel düşünme, problem çözme, iletişim kurma gibi becerilerin ölçülmesi, çoktan seçmeli testlerin uygulanmasının ötesinde yöntemlere ihtiyaç duyar. Bu ihtiyaçları performansın ölçülmesinin doğasında var olan özellikler karşılayabilmektedir. Matematik performansının ölçülmesi, öğrencilerin matematik performanslarını gösterebilecekleri görevlerden oluşur. Öğrencilerin bilgilerini bir araya getirmesine neden olacak fırsatlar sunarak daha kompleks düşünebilmelerini sağlamaya odaklanır. Matematik performansının ölçülmesi öğrencilerin öğrendikleriyle ilgili zengin bilgiler gerektirir. Matematik başarısının ölçülmesinde, öğrencilerin matematik performansının ölçülmesindeki görevlere ilişkin çalışmalarından elde edilen sonuçlar kadar sonuca ulaşmada sergiledikleri performans da önemlidir. Burada ölçmenin asıl odaklandığı nokta, doğru cevabın bulunmuş olması değil, öğrencilerin yaptıkları sentezi ortaya koyabilme derecesidir.

*"Matematik Eğitimi için Program ve Değerlendirme Standartları"*nın altını çizdiği ve Romberg ve diğerleri tarafından önerilen eğitim modelinde ele alındığı gibi performansın ölçülmesi "sürekli, dinamik ve genellikle resmi olmayan (informal) bir süreçtir" (Akt. Fine,1995). Performansın ölçülmesi, elde edilmek istenilen bilgiye bağlı olarak farklı yolların kullanılarak öğrencilerin öğrendikleriyle ilgili bilgi toplamayı, istenilen bu bilginin kullanımını ve bunlara ilişkin öğrencilerin düzeylerini geliştirmeyi içerir. Bu amaca hizmet eden ölçme araçlarından biri olarak performansın ölçülmesi matematik eğitimindeki önemini korumaktadır.

Performansın ölçülmesinde yer alan görevler, öğrencilerin matematiksel yetenekleri ve öğrendiklerine ilişkin çıkarımların geçerliğini arttırabilmektedir. Ancak performansın ölçülmesi bazı sınırlılıkları da beraberinde barındırmaktadır. Performansın ölçülmesiyle yeterli ölçmelerin yapılabilmesinde, gözlemleme ve raporlama sürecinde çok fazla zaman gerektirmektedir. Ayrıca ölçmenin güvenirliğini ve geçerliğini tehdit eden bazı durumlar da söz konusudur. Örneğin; performansın ölçülmesinde yer alan farklı görevler, güçlükleri açısından farklılık gösteriyorsa, öğrencinin puanı, sadece ölçmede yer alan bu görevlere bağlı olacaktır. Bu durumda, görevlerin değişkenliği öğrencinin matematiksel düşünme becerisindeki değişkenliğin potansiyel kaynağı olacaktır. Performansın ölçülmesinde puanlayıcılar da ölçmedeki bir başka değişkenlik kaynağıdır. Puanlayıcı, puanlamayı yapacak yeterli özelliğe sahip değilse öğrencilerin puanları seçilen bu puanlayıcıya bağlı olarak değişim gösterebilir (Fine, 1995). Böylece, performansın ölçülmesinde asıl gözlenmek istenen öğrenciler arası farklılıkların yanı sıra puanlayıcılardan kaynaklanan değişkenliklerin ortaya çıkması mümkün olabilmektedir. Genellikle, ölçülmek istenilen özellik açısından öğrencilerde var olan değişkenlik dışında farklı kaynaklardan oluşan değişkenlikler ölçmede istenmeyen değişkenlik kaynaklarıdır.

Görevler, maddeler, puanlayıcılar, zaman gibi değişkenlik kaynakları ve bunlar arasındaki etkileşim performansın ölçülmesiyle elde edilen sonuçların güvenirliğini etkileyebilmektedir. Bir ölçme aracının zamana karşı gösterdiği tutarlılığın derecesi güvenirlik olarak tanımlanabilir. Ölçme aracı ne kadar güvenilirse, bu ölçme aracının bir kez uygulanmasıyla elde edilen sonuçlar ile farklı zamanlarda tekrar uygulanmasıyla elde edilen sonuçların o derece tutarlı olması beklenir.

1.4. GÜVENİRLİK İLE GEÇERLİK ARASINDAKİ İLİŞKİ

Güvenirlik ile geçerlik arasındaki ilişki genel anlamda ele alınırsa, güvenirliğin geçerlik için bir önkoşul olduğu kabul edilmektedir (Gay, 1987). Bir başka deyişle, bir ölçmenin geçerli olabilmesi için önce güvenilir olması gerekir. Güvenirliği sağlamak, geçerliğin sağlanması için gerek şarttır ancak yeter şart değildir. Güvenilir bir değerlendirme her zaman geçerli olmayabilir. Bununla birlikte diğer bir açıdan bakıldığında, güvenirliğin geçerlik için bir ön koşul olmadığı görülebilir. Çünkü güvenirlik bazen geçerliği

olumsuz etkileyebilmektedir. Örneğin; yapılan tekrarlı ölçmeler sonucunda hep aynı sonuç elde ediliyorsa bu durum ölçmenin güvenilir olduğunu gösterir. Ancak yapılan bu ölçme işleminde ne ölçüldüğü bilinmiyorsa geçerlikten söz etmek mümkün değildir. Bunun dışında, bazen ölçülmek istenilen kapsam, güvenirliği arttırmak için daraltılır, homojen hale getirilir ki bu da geçerliğin düşmesine sebep olan bir diğer unsurdur. Sonuç olarak, daha güvenilir ölçümler elde etmek daha az geçerli ölçümler elde edilmesine sebep olabilmektedir (Baykul, 2000). Kısacası, hem güvenirlik hem de geçerlik, ölçmenin kalitesini belirleyen çok önemli faktörlerdir. Bir ölçmeden daima istenilen hem güvenilir hem de geçerli olmasıdır. Ancak bazen bu ikisinin birlikte elde edilmesi güç olabilir. Bu durumda ölçmenin türü ve amacı, ölçmeyi yapan kişinin hangisinin üzerinde daha çok durulması gerekli olduğuna karar vermesi konusunda önemli bir belirleyici olabilir.

1.5. PERFORMANSIN ÖLÇÜLMESİNDE GÜVENİRLİK

Güvenirlik, ölçme sonuçlarının tesadüfi hatalarından arınık olma derecesi olarak tanımlanmaktadır (Baykul, 2000). Güvenirlik, ölçümlerin zaman içersindeki tutarlılıklarının derecesidir. Bir test güvenilir olduğu ölçüde, bu testin farklı zamanlarda uygulanmasıyla elde edilen sonuçların benzer olma olasılığı o ölçüde artacaktır. Güvenirliğin derecesi genellikle bir katsayı ile ifade edilir. Bu katsayısı “0” (güvenilir değil) ile “1”(mükemmel güvenirlik) arasında değişen değerler alır ve ölçme sonuçlarının tesadüfi hatalardan (gelişi-güzel, random) ne derece arınık olduğunu gösterir. Ölçme sonuçları daima bir miktar hata barındırmaktadır, bir başka deyişle ölçme sonuçlarının güvenirliği daima bir miktar tehdit altında olmaktadır.

Performansın ölçülmesinde, güvenirlik en zayıf halka olarak düşünülmektedir (Alharby, 2006). Performansın ölçülmesinde puanlamanın puanlayıcıların kararlarına bağlı yapılması yani subjektifliği güvenirliği düşüren nedenlerden biridir. Puanlayıcı, performansın ölçülmesinin güvenirliğini düşüren önemli bir hata kaynağı olmakla birlikte, görev ya da zaman gibi diğer faktörlerle olan etkileşimi de en az o kadar önemli bir hata kaynağıdır. Bu nedenle sadece puanlayıcılar arası (interrater) ya da puanlayıcının kendi içindeki, puanlayıcı içi (intrarater) tutarlılığı sağlamak puanlamanın güvenirliği için yeterli olmamakta, hata kaynakları arasındaki etkileşimin de

güvenirliğin hesaplanmasında göz önünde bulundurulması gerekmektedir. Ancak güvenirliğin hesaplanması için kullanılan tüm yöntemler, hem puanlayıcı hem diğer hata kaynaklarıyla olan etkileşimi birlikte ele almamaktadır. Performansın ölçülmesinin güvenirliği, ölçmenin üç temel kuramı olan klasik test kuramı (KTK), madde tepki kuramı (MTK) ve genellenebilirlik kuramı (GK)'na dayalı yöntemlerle çalışılmaktadır. Güvenirliğin hesaplanmasında kullanılan bu üç kuram ve kuramlara dayalı yöntemler, üstün ve sınırlı yönleriyle birlikte aşağıda ayrıntılı olarak açıklanmıştır.

1.6. KLASİK TEST KURAMINA DAYALI YÖNTEMLER

Klasik test kuramının dayandığı temel varsayımlardan biri gözlenen puanın, gerçek puan ve hata puanı bileşenlerinin toplamından oluşmasıdır. Bu varsayım, X=T+E eşitliği ile gösterilir ve klasik test kuramının temel denklemi olarak ifade edilir (Baykul, 2000). Klasik test kuramının bir diğer varsayımı; aynı bireyin tekrarlı ölçmeler sonucu kestirilen gerçek puanları birbirinden doğrusal bağımsızdır. Bu varsayımlardan şu sonuçlara ulaşmak mümkündür: (a) Gözlenen puan varyansı, gerçek puan varyansı ve hata puanı varyansı toplamına eşittir. (b) Paralel formlardan elde edilen puanlar arasındaki korelasyon ya da parelel formlar arası güvenirlik gerçek puan varyansının gözlenen puan varyansı oranına eşittir. (c) Eğer güvenirliği R olan bir testin maddelerinin sayısı k oranında arttırılıyor ya da azaltılıyorsa, elde edilen yeni testin güvenirliği "k.R"ye (Spearman-Brown formülü adı verilir) eşit olacaktır (Keats, 1997).

Klasik test kuramında güvenirlik katsayısı, gerçek puan varyansının gözlenen puan varyansına oranı olarak tanımlanır. Buradaki gözlenen puan varyansı iki farklı varyans kaynağından oluşmaktadır: gerçek-puan varyansı (sistematik olduğu düşünülen) ve hata-puanı varyansı (gelişi-güzel (random) olduğu düşünülen). Gerçek varyans dışındaki tüm varyansın farklı hata kaynaklarından gelebileceği düşünülür. Genellikle ölçmeyle elde edilen puanların tutarlılığı olarak ifade edilen güvenirlik, hatanın bağlı olduğu kaynağa göre farklılık gösterebilir ve buna göre farklı isimlerle tanımlanır. Bir testin iki ya da daha fazla sayıdaki uygulaması (test-tekrar test) sonucu elde edilen puanlar arasındaki ilişki (korelasyon) "kararlılık" veya "tutarlılık" adını alır ve buradaki hata kaynağı uygulamalar arasındaki zaman olarak düşünülür. Fakat bu anlamdaki güvenirlikte madde örnekleminden kaynaklanan değişkenlik (varyans) göz önüne

alınmaz. Ölçme aracında bulunan maddelerden elde edilen puanların tutarlılığına "iç tutarlılık" anlamında güvenirlik denir ve buradaki hata kaynağı ölçme aracındaki her bir madde olmaktadır. Bu anlamdaki güvenirlikte ise zamana bağlı ortaya çıkabilecek değişkenlik göz önünde bulundurulmamaktadır. Eşdeğer (paralel) formlar güvenirliğinde aynı ölçme için iki paralel formun uygulanması sonucu elde edilen puanların "tutarlılığı" söz konusudur ve buradaki hata kaynağı ölçme aracı olarak değerlendirilir. Burada ise puanlayıcılardan doğacak değişkenlik söz konusu olmamaktadır. Ölçmede yer alan birden fazla puanlayıcının ölçülen aynı özelliğe verdikleri puanlar arasındaki tutarlılığa da "puanlayıcılar arası güvenirlik" denir ve bu kez buradaki hata kaynağı sadece puanlayıcılardır.

Performansın ölçülmesinde, önemli bir hata kaynağı olan puanlayıcılara bağlı olarak güvenirliğin hesaplanmasında; klasik test kuramı yaklaşımına dayalı farklı yöntemler bulunmaktadır. Bunları genel anlamda, puanlayıcı-içi (intrarater) güvenirlik ve puanlayıcılar-arası (interrater) güvenirlik hesaplama yöntemleri olarak ikiye ayırmak mümkündür (Alharby, 2006).

Puanlayıcı-içi (intrarater)güvenirlik: Bir puanlayıcının verdiği puanlardaki kendi kararlılığının göstergesidir. Bazı gelişi-güzel koşullar, örneğin; puanlayıcının yorgunluğu, psikolojik durumundaki değişmeler, puanlayıcının farklı zamanlarda kendi verdiği puanlar arasında uyumsuzluğun görülmesine sebep olabilir. Puanlayıcı-içi güvenirlik, bir puanlaycının farklı zamanlarda aynı bireylerin kağıtlarını birden fazla kez puanlaması ile elde edilen puanlar arası korelasyon katsayısının hesaplanmasıyla elde edilir.

Bu yöntem, test tekrar test güvenirliğin hesaplanmasına benzemektedir. Test tekrar test yönteminde, bir test bir gruba farklı zamanlarda iki kez uygulanır, farklı iki uygulamadan elde edilen puanlar arası korelasyon hesaplanarak, testin zamana karşı kararlığı anlamında güvenirliği kestirilir. Puanlayıcı-içi güvenirlikte ise bir kez uygulanan performansın ölçülmesinde bir puanlayıcının, iki farklı zamanda puanlamasıyla elde edilen puanlar arasındaki korelasyonun hesaplanmasıyla, puanlayıcının puanlamadaki tutarlılığı anlamında güvenirliğin kestirilmesi söz konusudur.

Ayrıca bir puanlayıcının aynı performansları iki kez puanlaması sonucunda elde edilen puanların uyumunun basit yüzdesi tutarlılık anlamında güvenirliğin bir ölçüsünü vermektedir. Basit yüzde; puanlayıcının aynı puanı verdiği bireylerin sayısının, puan verdiği tüm bireylerin sayısına oranı yüzdesidir. Örneğin; bir puanlayıcının 50 öğrencinin performansını iki farklı zamanda puanladığını düşünelim. Bu iki puanlama sonucunda 24 öğrencinin performansına her iki puanlamada aynı puanları vermiş 36 öğrencinin performansına ise farklı puanlar vermiş olsun. Bu durumda puanların uyumunun basit yüzdesi=100x(24/50)=%48 olacaktır. Basit yüzde yönteminin en önemli avantajı, hesaplanmasının kolay ve anlaşılır olmasıdır. Öte yandan, sınıflama, sıralama, aralık ve oran ölçeğinde elde edilen verilerin tümüne uygulanabilme özelliğine sahiptir (Goodwin, 2001). Basit yüzde hesaplanmasının en önemli sınırlılığı ise şansla ya da tesadüfen ortaya çıkan uyumu hesaplayamamasıdır.

Puanlayıcılar-arası (interrater) güvenirlik: Farklı puanlayıcıların birbirinden bağımsız olarak her bir öğrencinin kağıdına aynı puanı verme tutarlılıklarının derecesidir. Puanlayıcılar-arası güvenirlik en yaygın olarak, iki puanlayıcının performansa verdikleri puanlar arasındaki doğrusal ilişkinin derecesi olan korelasyon katsayısı (güvenirlik katsayısı olarak da bilinir) -Pearson çarpım moment korelasyon katsayısı (r)- ile hesaplanır. Bu korelasyon katsayısı; gerçek varyansın toplam varyansdaki yüzdesi olarak elde edilir. Burada geriye kalan yüzde puanlayıcıların puanları arasındaki uyumsuzluğu gösteren hata anlamı taşımaktadır. Örneğin; iki puanlayıcının puanları arasındaki korelasyon katsayısı 0.75 olarak bulunmuşsa, toplam varyansın %75'i gerçek varyans, kalan %25 ise bu iki puanlayıcının puanları arasındaki uyumsuzluğu gösteren hatayı ifade eder (Anastasi ve Urbina, 1997). Güvenirlik katsayısı olarak ele alınan Pearson çarpım moment korelasyon katsayının hesaplanması, çokça kullanılan, bilinen ve kolay yorumlanabilen bir teknik olarak avantaja sahiptir. Bunun yanı sıra korelasyon katsayısının hesaplandığı grubun büyüklüğünden doğrudan etkilenmesi ise bir dezavantaj olmaktadır (Goodwin, 2001).

Yine Pearson korelasyon katsayısının yalnızca iki değişken arasındaki doğrusal ilişkiyi hesaplamada kullanılabilir oluşu puanlayıcılar arası güvenirliğin hesaplanmasında başka bir sınırlılık olarak görülebilir. Pearson çarpım moment korelasyon katsayısının daha önemli bir sınırlılığı ise ortalamadan bağımsız oluşudur. Pearson çarpım moment

korelasyon katsayısı ortalamadan bağımsız olduğu için güvenirliğin hesaplanmasında iki puanlayıcıdan elde edilen puanlar arasındaki benzerlik ve farklılıkları göstermez. Birbirlerine göre puanlamada katılık/cömertlik açısından farklılık gösteren iki puanlayıcının verdikleri puanlar için hesaplanan Pearson çarpım moment korelasyon katsayısı pozitif ve bire yakın çıkabilir. Fakat bu durum bağımsız puanlayıcıların puanlarının birbirine benzer olduğunu değil, sadece puanlarının birlikte değiştiklerini gösterir. Puanlayıcıların verdikleri puanların ortalamaları arasında büyük farklılık olmasına rağmen puanlar arasındaki Pearson çarpım moment korelasyon katsayısı oldukça yüksek çıkabilir. Bu nedenle Pearson çarpım moment korelasyon katsayısı puanlayıcılar arası uyum ve güvenirliğin belirlenmesinde yetersiz kalmaktadır (Goodwin, 2001).

Puanlayıcılar arası güvenirliğin hesaplanmasında, korelasyon katsayısıyla birlikte puanların ortalamalarının karşılaştırılması uygun olmaktadır. Pearson çarpım moment korelasyon katsayısının güvenirlik kestirmek amacıyla kullanılmasında, sadece iki puanlayıcının bulunduğu durumda puanlayıcılardan elde edilen puan ortalamalarının eşleştirilmiş gruplar için t testi, ikiden fazla puanlayıcının bulunduğu durumda ise tekrarlı ölçmeler için ANOVA kullanılarak karşılaştırılması ve hesaplanan korelasyon katsayısıyla beraber yorumlanması daha doğru olabilir (Goodwin, 2001).

Puanlayıcılar-arası güvenirliği bulmanın bir başka yolu, puanlayıcılar arası kesin uyumun yüzdesinin hesaplanmasıdır. Bu hesaplama, puanlayıcılar tarafından aynı puanı alan öğrencilerin sayısının, ölçmeye katılan tüm öğrencilerin sayısına bölünmesiyle yapılmaktadır (Meyer, 1999). Puanlar arası uyumun basit yüzdesinin heaplanması yöntemiyle, puanlayıcılar arasındaki uyumsuzlukların belirlenmesi mümkün olur. Böylece, puanlayıcıların puanlamaya ilişkin eğitilmesi gerekliliği ya da eğitilmiş puanlayıcılara verilen eğitimin yetersizlikleri kolayca belirlenebilir (Wolf, 1997). Bu yöntemin bir diğer avantajı, tüm ölçek türlerinden elde edilen verilere (sınıflama, sıralama, eşit aralıklı ve eşit oranlı) uygulanabilir olmasıdır. Aynı zamanda, uyumun basit yüzdesini hesaplamak ve anlamak oldukça kolaydır (Goodwin, 2001). Ancak basit yüzde hesaplanmasının en önemli sınırlılığı, şansla ya da tesadüfen ortaya çıkan uyumu hesaplayamamasıdır. Cohen'in Kappa'sı (1960), puanlayıcılar arası şansa bağlı ortaya çıkabilecek olası uyumun düzeltilmesini içeren güvenirlik katsayısının hesaplanmasına

bağlı bir yöntem olarak, basit yüzde yönteminin sınırlılığına bir çözüm sağlamaktadır. Kappa istatistiği, şans başarısı üzerindeki uyumun kestirimini sağlar. Puanlayıcılar arasındaki uyum, şansla ortaya çıkabilecek uyum düzeyinde ise kappa sıfıra eşit çıkar. Eğer uyum şansla ortaya çıkabilecek uyum düzeyinin üzerinde ise kappa sıfırdan daha büyük bir değer olacaktır. Kappa eşitliği;

$$Kappa = \frac{P_0 - P_c}{1 - P_c}$$

denklemi ile ifade edilir. Bu eşitlikte yer alan P_0 gözlenen uyum oranını, P_c şansla ortaya çıkabilecek uyumun oranını ifade eder. Kappa istatistiği teorik olarak -1 ile +1 arasında değerler almakla birlikte, sıfırdan küçük değerler şansla ortaya çıkması beklenen uyumun altında uyum düzeyini gösterdikleri için negatif değerler dikkate alınmamaktadır (Goodwin, 2001). Kappa istatistiği, tüm uyumsuzlukları eşit düzeyde ele almaktadır. Bu sınırlılığı ortadan kaldırmak üzere, Cohen ağırlıklandırılmış kappa (weighted kappa) istatistiğini geliştirmiştir. Ağırlıklandırılmış kappa istatistiğinde, bazı uyumsuzluklara hiç ağırlık verilmezken, bazı uyumsuzluklara kısmi ağırlıklandırma yapılır. Ağırlıklandırılmış kappa istatistiği, kappa istatistiği gibi şansla ortaya çıkabilecek uyuma göre düzeltmeyi içermektedir. Ancak bu istatistiklerin bazı sınırlılıkları da bulunmaktadır. İki puanlayıcının puanlarının marjinal (uçlardaki) dağılımları aynı olduğunda kappa ve ağırlıklandırılmış kappa katsayıları teorik olarak mümkün olan en yüksek değer 1'e eşit olabilirler. Diğer bir sınırlılık ise daha az ya da daha çok madde veya puanlayıcı olması hallerinde, güvenirliğin ne olacağının kestiriminde kullanılan Sperman-Brown formülü gibi bir formülün, kappa ve ağırlıklandırılmış kappa için kullanılmasının mümkün olmamasıdır (Goodwin, 2001). Puanlayıcılar arası uyumu ya da güvenirliği tespit etmede Cohen'in kappa istatistiğinin bir diğer sınırlılığı ise puanlayıcılar dışında başka değişkenlik kaynaklarının olması durumunda, bu kaynaklardan meydana gelebilecek potansiyel hataların güvenirlik hesaplanmasında göz önüne alınamamasıdır (Goodwin, 2001). İki puanlayıcı arasındaki uyumun hesaplanmasını sağlayan Cohen'in kappasını Fleiss (1971), ikiden daha fazla puanlayıcının puanlama yaptığı durumlara genişletmiştir. Fleiss'in kappası, tüm puanlayıcıların rasgele uyumlarının üzerinde gözlenen uyumlarının derecesini verir ve aşağıdaki eşitlik ile hesaplanır:

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}$$

Bu denklemde $1 - \bar{P}_e$ şansla ortaya çıkabilecek uyumun büyüklüğünü, $\bar{P} - \bar{P}_e$ değeri ise gözlenen uyumun büyüklüğünü vermektedir. Fleiss'in kappası 0 ile 1 arasında değişen değerler alır ve 1 değeri puanlayıcılar arası mükemmel uyumu ifade eder (Fleiss, 1971).

İkiden fazla puanlayıcının belirli bir davranışı, performansı, soruyu vb. puanlaması durumunda, puanlayıcılar arası güvenirliğin belirlenmesinde kullanılan bir diğer yöntem, parametrik olmayan bir istatistiksel teknik olan Kendall'ın uyum katsayısıdır (Howell, 2002) ve aşağıda gösterildiği şekilde hesaplanır:

$$W = \frac{12\sum T_j^2}{k^2 N(N^2 - 1)} \qquad \frac{3(N+1)}{N-1}$$

Bu eşitlikte;

k: puanlayıcıların sayısı,

N: puanlanan görev ya da madde sayısı,

T_j: Her bir madde ya da göreve tüm puanlayıcıların verdiği puanların toplamını göstermektedir.

Kendall'ın uyum katsayısına ilişkin yokluk hipoteziyle de yorumlamaya gidilebilir. Puanlayıcılar arası uyumun olmadığı yokluk hipotezi puanlayıcı sayısının yediye eşit ve büyük olduğu durumlar için X^2 ile test edilebilir ve $X^2_{(N-1)}$=k (N-1)W değeri N-1 serbestlik derecesinde yaklaşık X^2 dağılımı gösterir. Ancak genellikle puanlayıcı sayısı, manidarlık testi için gerekli olan minimum sayı olan yediden az olduğu için, bu test nadiren uygulanır (Howell, 2002; Cooper, 1997).

Performansın ölçülmesinde puanlayıcılar önemli bir hata kaynağı olmasına karşın aynı zamanda farklı hata kaynakları ve bunlar arasındaki etkileşimden doğan hata kaynakları da bulunmaktadır. Ancak yukarıda belirtilen yöntemlerle güvenirliğin hesaplanmasında, birden fazla hata kaynağı ve bunlar arasındaki etkileşim birlikte göz önünde bulundurulamaz. Böylece klasik test kuramında, farklı varyans kaynaklarının

büyüklüğünün kestirilmesi çoklu analizlerin yapılmasını gerektirdiği gibi, farklı varyans kaynakları arasındaki etkileşimin kestirilmesi mümkün olmamaktadır. Ayrıca klasik güvenirlik kuramı, sadece ölçme yapılan bireylerin birbirlerine göre durumlarına ilişkin kararlar çerçevesinde (göreceli değerlendirme) güvenirliğin hesaplanmasına olanak sağlar, öğrencilerin mutlak değerlendirilmesine ilişkin güvenirliğin hesaplanmasına imkan tanımaz (Fine, 1995). Özetle, klasik test kuramı yaklaşımları performansın ölçülmesinde güvenirliğinin hesaplanmasında yetersiz kalmaktadır.

Klasik test kuramının aksine, genellenebilirlik kuramına dayalı yapılan güvenirlik çalışmalarında; puanlayıcılardan, diğer değişkenlik kaynaklarından ve bunlar arası etkileşimden ortaya çıkan hata varyansları bir arada göz önüne alınabilmektedir.

1.7. GENELLENEBİLİRLİK KURAMI

Genellenebilirlik (G) kuramı ölçme sonuçlarının güvenirliğinin belirlenmesini, güvenilir gözlemlerin tasarımını, araştırılmasını ve kavramsallaştırılmasını sağlayan istatistiksel bir kuramdır. G kuramı, klasik test kuramının bir uzantısıdır (Cronbach ve diğerleri, 1972; Brennan, 2001). Klasik test kuramı, bir tek gerçek puana sahip her bir gözlem ya da test puanının, paralel gözlemlerin bir grubuna ait tek bir güvenirlik katsayısı üretmesi varsayımı etrafında merkezlenir. Bu varsayım, paralel formlar dikkatli bir şekilde eşitlendiğinde makulken puanların ortalamaları ve varyansları farklı ya da formlardaki maddeler heterojen olduğunda gerçekçi olmaz ve oldukça kısıtlayıcıdır.

Çok boyutlu bir ölçmede iç tutarlılık anlamında güvenirlik düşük olma eğilimindeyken aynı zamanda test tekrar test ve paralel formlar güvenirliği yüksek olabilmektedir. Bu durum klasik güvenirlik modelinin sınırlılık ve çelişkilerini gösterir. Bu sınırlılığın temel nedeni; ölçme sonuçlarına karışan hataların, klasik kuramda sadece bir değişkenlik kaynağından gelen hatalar olarak ele alınmasındadır. Nitekim klasik test kuramının güvenirlik hesaplama yöntemleri ele aldığı tek bir hata kaynağına göre farklılık göstermektedir. Bu farklılıklar aşağıda belirtilmiştir:

- Test tekrar test yöntemiyle güvenirlik hesaplanırken zamana bağlı *kararlılık* anlamında güvenirlik söz konusudur ve burada hata kaynağı olarak *zaman* ele alınır.

- Paralel formlar yöntemiyle güvenirlik hesaplanırken bir ölçme aracının paralel formları arasındaki *tutarlılık* anlamında güvenirlik söz konusudur ve buradaki hata kaynağı *formlar* olmaktadır.
- İç tutarlılık anlamında güvenirliğin hesaplanmasında kullanılan yöntemlerde ise ölçme aracını oluşturan görev ya da maddelerin kendi içindeki tutarlılıkları anlamında güvenirlikten söz edilir ve burada hata kaynağı olarak *görevler* ya da *maddeler* dikkate alınır.

Buradaki ifadelerden anlaşılacağı üzere, aynı ölçmeye ilişkin olarak klasik test kuramının farklı hata kaynaklarını dikkate alan farklı yöntemlerle hesaplanan güvenirlik katsayıları birbirinden farklı olabilir. G kuramı, klasik test kuramının bu sınırlılığını ortadan kaldırmaya yönelik daha esnek bir yöntem olarak geliştirilmiş olup puanlayıcı, zaman, ölçme formu, görevler ya da maddeler gibi tüm potansiyel değişkenlik kaynaklarından gelebilecek hataları birlikte ve aynı zamanda değerlendirerek kapsamlı tek bir güvenirlik katsayısı hesaplanmasına imkan vermektedir.

Klasik test kuramında güvenirliğin hesaplanması yoluyla, örneklemden elde edilen verinin güvenirliğine ulaşılır. Oysa G kuramı, bir gözlemin güvenirliğinin sonuç çıkarılmak istenen evrene bağlı olduğunu kabul eder. G kuramı, tesadüfi olarak seçilmiş bir örneklemden, örneklemin seçildiği evrene doğrulukla genellenebilmesine ilişkin bir kuram olarak klasik test kuramını yeniden yorumlamaktadır.

Genellenebilirlik kuramında öğrencinin bir performansın ölçülmesinde yer alan görevde gösterdiği performansı, puanlaycı ve görev gibi olası tüm değişkenlik kaynaklarının bir arada bulunduğu karmaşık bir evrendeki performansından çekilen bir örneklemi olarak görülür. Shavelson ve Webb (1991)'e göre, G kuramı dört farklı açıdan klasik test kuramının daha genişletilmiş bir halidir: 1. Genellenebilirlik kuramı, çoklu varyans kaynaklarını tek bir analizde ele alır. 2. Her bir varyans kaynağının büyüklüğünün belirlenmesini sağlar. 3. Hem bireylerin performanslarına dayalı göreceli kararlar hem de bireylerin performanslarıyla ilgili mutlak kararlar alınmasına ilişkin iki farklı güvenirlik katsayısının hesaplanmasına olanak tanır. 4. Belirli bir amaca bağlı olarak, ölçme hatasının en aza indirgenebileceği ölçmelerin düzenlenmesine (D-çalışmaları)

imkan tanır. Kısacası Genellenebilirlik kuramı performansın ölçülmesinde güvenirliğin kestirilmesine uygun bir kuramdır.

Genellenebilirlik kuramı, klasik test kuramının sadece yeniden yorumlanması olmayıp, aynı zamanda güvenirlik ile geçerlik arasındaki yerleşe gelmiş farkın, güvenilir ölçmeler düzenleyerek nasıl ortadan kaldırılabileceğini de ortaya koymaktadır. Genel anlamıyla yapı geçerliği, ölçülmek istenilen bir yapının yani bireylerde var olduğu kabul edilen bir özelliğin, ölçme sonucunda ortaya konulma derecesi olarak yorumlanabilir (Baykul, 2000). Oysa klasik kurama göre güvenirlik, paralel ölçmeler sonucunda gerçek puana ilişkin doğru tahminde bulunmanın bir derecesidir. G kuramında yer alan "evren" kavramı; tüm gözlem koşulları ve değişkenlik kaynaklarını kapsamaktadır ki bu da klasik kuramdaki geçerlikte yer alan "yapı" kavramını tanımlamaktadır. G kuramı, bu örtük yapıya (kabul edilebilir gözlemlerin evreni) ilişkin kestirimlerin doğru olarak elde edilmesini sağlarsa, güvenilir sonuçlara ulaşılabileceğini ifade eder. Böylelikle G kuramı, güvenirlik ile geçerlik arasındaki geleneksel ayrımı ortadan kaldırır (Allal ve Cardinet, 1997).

Genellenebilirlik kuramını açıklarken bazı kavram ve terminolojilerin açıklanması konunun daha iyi anlaşılmasında faydalı olacaktır. Genellenebilirlik kuramında, madde ya da görev, zaman, puanlayıcı gibi ölçmenin benzer durumlarının setine, *değişkenlik kaynağı* (*facet*) (Brennan, 2001) adı verilir. Bir başka deyişle değişkenlik kaynağı, ölçme hatasının olası kaynağı olarak tanımlanabilir. Böylece değişkenlik kaynağıyla ilişkili bir varyans istenen bir varyans olmayıp, bu tür bir varyansın olabildiğince az olması istenen bir durumdur (Alharby, 2006). Değişkenlik kaynaklarının düzeyleri *koşullar (conditions)* olarak adlandırılır. Örneğin; "puanlayıcılar" bir değişkenlik kaynağı ise birinci puanlayıcı, ikinci puanlayıcı, üçüncü puanlayıcı vb.nin her biri bir koşuldur. Genellikle var olan bir değişkenlik kaynağının olası koşullarının sonsuz büyüklükte olduğu varsayılır. Gözlemlerin yapıldığı eldeki örneklemin yerine geçebilecek olası gözlemlerin tümüne "*kabul edilebilir gözlemlerin evreni (the universe of admissible observation)"* denir. "*Genelleme evreni (the universe of generalization)"* ise araştırmacının genellemek istediği koşulların setidir. Pekçok ölçme durumunda bireyler, *ölçmenin amacı (the object of measurement)*dırlar. Bir başka deyişle bireyler, istenilen kararların alınacağı ölçmenin hedefleri durumundadırlar. Bu sebeple bireylere

bağlı varyans istenilen bir durum olduğu için bireyler değişkenlik kaynağı olarak düşünülmez. *Evren puanı* amaçlanan değişkenlik kaynakları için kabul edilebilir gözlemlerin evreninden elde edilecek puanların ortalaması olarak bulunan, ölçme puanı olarak tanımlanır. Evren puanları varyansı, klasik test kuramında yer alan gerçek puan varyansına benzer. Klasik test kuramından farklı olarak, G kuramında iki ayrı hata varyansı bulunur. Bu farklılık, G kuramında iki ayrı anlamda karar vermenin mümkün olmasından kaynaklanmaktadır. G kuramında hem göreceli hem de mutlak hata varyansı bulunmaktadır. G kuramındaki göreceli hata varyansı klasik kuramdaki hata varyansı gibi düşünülebilir (Shavelson ve Webb, 1991). Kısaca klasik test kuramında;

Güvenirlik katsayısı = Gerçek puan varyansı / (Gerçek puan varyansı + Hata puanı varyansı),

Genellenebilirlik kuramında;

Genellenebilirlik katsayısı = Evren puan varyansı / (Evren puan varyansı + Göreceli hata puanı varyansı)

olarak ifade edilebilir.

Yukarıdaki temel kavram ve terimlere ek olarak genellenebilirlik kuramında farklı desen ve yöntemler de bulunmaktadır: Çalışmanın amacına bağlı olarak; G ve D çalışmaları, değişkenlik kaynağının tesadüfi (random) ya da sabit (fixed) olması, verilerin toplanmasında desenin çapraz (cross) ya da yuvalanmış (nested) olması, sonuçların yorumlanmasının göreceli ya da mutlak anlamda yapılıyor olması ve genellenebilirlik (G) katsayısının ya da güvenirlik (Phi) katsayısının hesaplanmasının mümkün olması. Bunların her biri aşağıda kısaca açıklanmıştır.

G-(Study) Çalışmaları ve D-(Study) Çalışmaları: G Kuramında, güvenirliğin araştırılmasında iki aşamadan söz etmek mümkündür (Goodwin, 2001). Bunlardan ilki *Genellenebilirlik çalışması (G-çalışması)* ve ikincisi *Karar çalışması (D-çalışması)*dır.

G-çalışması sürecinde, örneklemin evrene genellenebilmesi için, puanların değişkenliğinin tüm kaynakları (varyans bileşenleri) ve bunlar arasındaki etkileşimler aynı anda ANOVA yöntemi kullanılarak kestirilir. Kestirilen bu varyans bileşenleri bir sonraki aşama olan D-çalışmasında kullanılır. *D-çalışması*, karar vermek üzere belirli

bir amaç için veri toplanan çalışmadır. Yapılan bir D çalışmasında, incelenen bireyleri tanımlamak için veri toplanabilir. Örneğin, öğrenciler arasında seçme yapmak ya da öğrencileri bir programa yerleştirmek, bir sınavdaki grupları karşılaştırmak ya da iki ya da daha fazla değişken arasındaki ilişkiyi araştırmak için D-çalışması düzenlenir. Aslında D-çalışması "Eğer ki... ne olur?" sorusuna cevap aramak üzere yapılan bir çalışmadır (Alharby, 2006). D-çalışması sürecinde araştırmacı, farklı senaryolar kullanarak daha yüksek güvenirlik ve daha düşük hataların olduğu durumlar elde etmeye çalışır. Örneğin; "Daha fazla sayıda puanlayıcı kullandığımda ne olur?", "Görev sayısını azaltırsam ölçme aracımın güvenirliğinde kaybım çok fazla olur mu?" vb. sorularının cevabı D-çalışması yapılarak araştırılır. Sonuç olarak, tek bir G-çalışmasından elde edilen aynı varyans kestirimlerine dayalı pek çok D-çalışması düzenlenebilir. Bir başka deyişle ölçme durumlarında ölçme aracında yer alan maddelerin, ölçmenin uygulanma zamanlarının ya da ölçmeyi yapan puanlayıcıların sayılarının iki, üç... katına çıkarıldığında ya da belli oranlarda azaltıldığında ne olacağını kestirmede D-çalışması kullanılabilir. D-çalışmasında kullanılan formül Spermen-Brown formülüne benzerdir (Mushquash ve O'Connor, 2006).

Klasik test kuramında gerçek puan kavramı, bireyin gerçek puanı; tam anlamıyla paralel olan çok sayıdaki ölçme sonuçlarının ortalaması olarak tanımlanır. Gerçek puan varyansı da, bu ortalamanın varyansıdır. Güvenirlik, gerçek puan varyansının gözlenen puan varyansına oranı olarak ifade edilir. Genellenebililik kuramında ise bireyin evren puanı, genellenebilirlik evrenindeki ölçmelerin ortalaması olarak tanımlanır. Klasik test kuramında yer alan bu ölçmelerin paralel olduğu varsayımı genellenebilirlik kuramında yoktur. Genellenebilirlik katsayısı, evren puan varyansının gözlenen puan varyansına oranıdır. Genellenebilirlik katsayısı D-çalışması desenine bağlıdır. Bir başka deyişle, farklı D-çalışması desenleri için farklı genellenebilirlik katsayısı elde edilebilir. Genellenebilirlik katsayısının başka bir tanımı, evren ile gözlenen puanlar arasındaki korelasyonun karesidir. Evren puanı ve evren puan varyansı ise genellenebilirlik evrenine bağlıdır. Böylece, eğer iki araştırmacı aynı D-çalışma desenini farklı genellenebilirlik evreni üzerinden yapıyorsa farklı genellenebilirlik katsayıları elde edeceklerdir.

Tesadüfi Değişkenlik Kaynağı ve Sabit Değişkenlik Kaynağı: Bir değişkenlik kaynağının tesadüfi olarak mı yoksa sabit olarak mı ele alınacağı tamamıyla araştırmacının kararına bağlıdır. Eğer araştırmacı, örneklemin ötesinde genelleme evrenine genelleme yapmak istiyorsa, bu durumda değişkenlik kaynağı tesadüfi olarak ele alınacaktır. Diğer taraftan, eğer araştırmacı örneklemin ötesinde bir genellemeye gitmek istemiyorsa, o zamanda değişkenlik kaynağı sabit olarak düşünülecektir. Örneğin; bir matematik performansının ölçülmesinde öğrencilere iki farklı görev verildiği ve iki farklı puanlayıcının her iki görevi de bağımsız olarak puanlayacağı varsayılsın. Eğer araştırmacı "Evrendeki tüm olası puanlayıcılara genelleme yapmak istiyorum" diyorsa, puanlayıcı değişkenlik kaynağı, tesadüfi olacaktır. Eğer araştırmacı "Öğrencilerin sadece bu iki görevdeki durumunun nasıl olduğunu belirlemek istiyorum" diyorsa, bu durumda görev değişkenlik kaynağı sabit olacaktır (Alharby, 2006). Değişkenlik kaynağının tesadüfi ya da sabit olarak ele alınması güvenirliğin kestirimini etkilemektedir. Bu duruma daha ileriki bir bölümde yeniden değinilecektir.

Çaprazlanmış Desen ve Yuvalanmış Desen: Bu iki kavramı açıklamak için yukarıdaki örnekten (matematik performansının ölçülmesi) yararlanılacaktır. Eğer her bir öğrenci (b) her bir görevi (g) cevaplandırır ve her bir puanlayıcı (p) da her bir öğrencinin her görevini puanlarsa, bu desen tümüyle çaprazlanmış bir desen olacaktır. Bu desen "b x g x p" şeklinde gösterilir. Çalışmalarda tümüyle çaprazlanmış desenler tercih edilmektedir. Çünkü bu tür desenler mümkün olan tüm değişkenlik kaynaklarından ve bunlar arasındaki etkileşimlerden ortaya çıkan hataların kestirimlerine olanak sağlar (Mushquash ve O'Connor, 2006). Çaprazlanmış desenlerin dışında, eğer farklı öğrencilere (b), farklı görevler (g) sunulursa ve farklı puanlayıcılar (p) farklı öğrenciler tarafından cevaplanan farklı görevleri puanlarsa, bu desenlere de tümüyle yuvalanmış desen adı verilir. Bu desende puanlayıcılar görevlerle, görevler de öğrencilerle yuvalanmıştır ve bu desen "p : g : b" olarak ifade edilir. Bunlara ek olarak, bazı değişkenlik kaynaklarının çaprazlanmış bazılarının ise yuvalanmış olduğu karışık desenler de düzenlenebilir.

Göreceli Model ve Mutlak Model: Genellenebilirlik kuramı çerçevesinde, hangi modelin uygun olduğu daima uygulama amacına yani araştırmacının ölçme sonuçlarını nasıl kullanacağına bağlıdır. Eğer öğrenciler birbiriyle kıyaslanacaksa, her bir

öğrencinin bir göreve ilişkin göstermiş olduğu performansın puanlanması, içinde bulunduğu gruptaki diğer öğrencilerin nasıl performans gösterdiklerine bağlı olarak yapılacaktır (bu öğrencinin grubun dağılımındaki yeri neresidir ?). Bu durumda araştırmacı verilerin analizi için göreceli modeli kullanacaktır. Göreceli genellenebilirlik katsayısı, öğrencilerin ham puanlarının büyüklüğündeki olası değişimlere bakılmaksızın, değişkenlik kaynaklarındaki sıralamadaki yerlerini koruma derecelerini göstermektedir. Göreceli genellenebilirlik katsayısı, klasik test kuramındaki güvenirlik katsayısına karşılık gelmektedir (Mushquash ve O'Connor, 2006).

Performansın ölçülmesindeki diğer bir olasılık, araştırmacının öğrencileri ölçerek mutlak bir kritere (ölçüt) bağlı olarak değerlendirme yapmasıdır. Bu durumda öğrencilerin performansı içinde bulundukları grubun performansı dikkate alınmaksızın, önceden belirlenmiş bir kritere göre değerlendirilir. Bu durumda araştırmacı mutlak modeli kullanır. Mutlak genellenebilirlik katsayısı, göreceliden daha katıdır ve hem ölçülen öğrencilerin sıralamadaki yerlerinin düzenindeki uyumun derecesini hem de ham puanlarının büyüklüklerindeki uyumun derecesini yansıtır. Mutlak genellenebilirlik katsayısı, araştırmacı için elde edilen puanların gerçek değerlerinin önemli ya da daha anlamlı olduğunda kullanışlı olabilecek bir katsayıdır (Mushquash ve O'Connor, 2006).

Genellenebilirlik Katsayısı (G-katsayısı) ve Güvenirlik İndeksi (Phi-katsayısı): Genellenebilirlik katsayısı (G-katsayısı), klasik test kuramındaki gerçek varyansın gözlenen varyansa oranı olan, güvenirlik katsayısıyla benzerlik gösterir. Özellikle (1) değişkenlik kaynağının sadece maddeler olduğu, (2) D çalışmasındaki düzeylerin testte yer alan maddelerin sayısına eşit olduğu ve (3) sadece göreceli modelin kullanıldığı durumda, tek değişkenlik kaynaklı (maddeler) çaprazlanmış desenlerdeki G-katsayısı, klasik test kuramındaki Cronbach alfa'nın eşiti olmaktadır (Sudweeks, Reeve ve Bradshaw, 2005). G-katsayısı, göreceli modellerde kullanılır ve gerçek varyansın, gerçek varyans ile göreceli varyansın toplamına bölünmesiyle hesaplanır. Güvenirlik indeksi (Phi-katsayısı) ise mutlak modeller için kullanılmaktadır. Güvenirlik indeksi gerçek varyansın, gerçek varyans ile mutlak varyansın toplamına bölünmesiyle elde edilir. Diğer bir deyişle, bu iki katsayı hatanın nasıl ele alındığına bağlı olarak değişmektedir.

Değişkenlik Kaynağı olarak Puanlayıcılar: "Puanlayıcılar"ın genellenebilirlik kuramında bir değişkenlik kaynağı olarak nasıl ele alınacağı hipotetik bir örnekle açıklanmaya çalışılacaktır. Bu örneğe dayanarak, değişkenlik kaynağı olarak ele alınan "puanlayıcılar"ın durumundaki değişiklik, ilgili hata terimlerinin de değişmesine sebep olacaktır. Puanlayıcıların ve görevlerin iki değişkenlik kaynağını oluşturduğu, öğrencilerin ise ölçmenin amacı olduğu bir matematik performansının ölçülmesi ele alınacaktır. Buradan bir G-çalışması ve farklı D-çalışması senaryoları oluşturulacaktır. Buradaki G-çalışması iki değişkenlik kaynaklı tamamıyla çaprazlanmış (b x g x p) bir desen üzerinden yapılmaktadır. Bu durumda, toplam gözlenen varyans aşağıda ifade edilen yedi bağımsız varyans bileşeninden oluşur (Alharby, 2006):

$$\sigma^2_{(X_{bgp})}=\sigma^2_{(b)}+\sigma^2_{(g)}+\sigma^2_{(p)}+\sigma^2_{(bg)}+\sigma^2_{(bp)}+\sigma^2_{(gp)}+\sigma^2_{(bgp)} \quad (1)$$

Her bir varyans bileşeni, varyans analizi ile aşağıda verilen Tablo 3 yardımıyla kestirilebilir:

Tablo 3. İki Değişkenlik Kaynaklı Tesadüfi Desen İçin Varyans Bileşenlerinin Kestirilmesi

Varyans Kaynağı	Kareler Toplamı	Sd	Kareler Ortalaması	Kestirilen Varyans Bileşenleri
Birey (b)	SS_b	n_b-1	$MS_b= SS_b/ n_b$-1	$\sigma^2_{(b)}$
Puanlayıcı (p)	SS_p	n_p-1	$MS_p= SS_p/ n_p$-1	$\sigma^2_{(p)}$
Görev (g)	SS_g	n_g-1	$MS_g= SS_g/ n_g$-1	$\sigma^2_{(g)}$
b x p	SS_{bp}	$(n_b$-1$)(n_p$-1$)$	$MS_{bp}= SS_{bp}/ n_{bp}$-1	$\sigma^2_{(bp)}$
b x g	SS_{bg}	$(n_b$-1$)(n_g$-1$)$	$MS_{bg}= SS_{bg}/ n_{bg}$-1	$\sigma^2_{(bg)}$
p x g	SS_{pg}	$(n_p$-1$)(n_g$-1$)$	$MS_{pg}= SS_{pg}/ n_{pg}$-1	$\sigma^2_{(gp)}$
b x p x g, e	$SS_{bpg,e}$	$(n_b$-1$)(n_p$-1$)(n_g$-1$)$	$MS_{bpg,e}= SS_{bpg,e}/ n_{bpg,e}$-1	$\sigma^2_{(bgp)}$

Gruplar arası farklılığın istatistiksel olarak manidar olup olmadığının test edilmesi için varyans bileşenlerinden F değerlerinin hesaplanması, G çalışmasında yapılmamaktadır (Shavelson ve Webb, 1991; Brennan, 2001). Tablo 3'ün son sütununda verilen varyans bileşenleri G-çalışmasında, aşağıda gösterildiği şekilde elde edilebilir:

Bu G-çalışmasına bağlı farklı farklı D-çalışmları oluşturulabilir. Burada sadece, puanlayıcıların bir değişkenlik kaynağı olarak ele alındığı üç farklı D-çalışması açıklanmaya çalışılmıştır (Alharby, 2006):

1. D-Çalışması: İki görevin ve iki puanlayıcının bulunduğu, çaprazlanmış desen ve göreceli model düzenlenmiş olsun. Bu durumda göreceli varyans σ^2:

$$\sigma^2_{(\delta)} = \frac{\sigma^2_{bg}}{N_g} + \frac{\sigma^2_{bp}}{N_p} + \frac{\sigma^2_{bgp}}{N_g N_p} = \frac{\sigma^2_{bg}}{2} + \frac{\sigma^2_{bp}}{2} + \frac{\sigma^2_{bgp}}{2x2} \tag{3}$$

olacaktır. Burada σ^2_b gerçek varyanstır ve σ^2_g, σ^2_p ve σ^2_{gp} ise etkileşimsiz varyanslar olarak düşünülürler ve model göreceli olduğu için eşitlikte yer almazlar. Bu durumdaki G-katsayısı, gerçek varyansın, gerçek varyans ile etkileşimli varyansların toplamına bölünmesiyle hesaplanır:

$$G - katsayısı = \frac{\sigma^2_b}{\sigma^2_b + \frac{\sigma^2_{bg}}{2} + \frac{\sigma^2_{bp}}{2} + \frac{\sigma^2_{bgp}}{2x2}} \tag{4}$$

Şimdi de varsayalım ki aynı durum için göreceli model yerine mutlak model kullanılıyor olsun. Görev, puanlayıcı ve bunlar arasındaki etkileşime bağlı varyans bileşenleri mutlak varyansın parçaları olarak ele alınır ve önceki hesaplamada yer alan paydaya eklenerek daha büyük bir payda elde edilir. Bu da elde edilecek katsayının daha küçülmesine sebep olur. Bu durumda, güvenirlik katsayısı denilen Phi katsayısı;

$$\Phi - katsayısı = \frac{\sigma^2_b}{\sigma^2_b + \frac{\sigma^2_g}{N_g} + \frac{\sigma^2_p}{N_p} + \frac{\sigma^2_{bg}}{N_g} + \frac{\sigma^2_{bp}}{N_p} + \frac{\sigma^2_{gp}}{N_g N_p} + \frac{\sigma^2_{bgp}}{N_g N_p}}$$

$$= \frac{\sigma^2_b}{\sigma^2_b + \frac{\sigma^2_g}{2} + \frac{\sigma^2_p}{2} + \frac{\sigma^2_{bg}}{2} + \frac{\sigma^2_{bp}}{2} + \frac{\sigma^2_{gp}}{2x2} + \frac{\sigma^2_{bgp}}{2x2}} \tag{5}$$

şeklinde elde edilir.

(5) ve (6) eşitliklerinden puanlayıcıların sayısının arttırılmasının ilgili hatayı düşüreceği ve sonuç olarak da güvenirliği arttıracağı açıkça görülmektedir. Bu demektir ki,

ölçmeye daha fazla sayıda puanlayıcı eklemek, genellenebilirlik kuramı çerçevesinde güvenirlik katsayısını arttırmaktadır.

2. D-Çalışması: Bu kez yine aynı sayıda görev (2) ve puanlayıcıdan (2) oluşan, g x (p : b) olarak gösterilen bir karışık desen durumunu ele alalım. Burada, öğrenciler görevlerle çaprazlanmış, ancak puanlayıcılar ile öğrenciler yuvalanmıştır. Bu demektir ki; her bir öğrenci her bir görevi yerine getirirken, tüm görevleri gerçekleştiren farklı öğrencilerin performanslarını farklı puanlayıcılar puanlamaktadır. Eğer modelin göreceli olmasına karar verilmişse, o zaman göreceli hata varyansı:

$$\sigma^2_{(\delta)} = \frac{\sigma^2_p}{N_p} + \frac{\sigma^2_{bp}}{N_p} + \frac{\sigma^2_{bg}}{N_g} + \frac{\sigma^2_{gp}}{N_p N_g} + \frac{\sigma^2_{bgp}}{N_p N_g} = \frac{\sigma^2_p}{2} + \frac{\sigma^2_{bp}}{2} + \frac{\sigma^2_{bg}}{2} + \frac{\sigma^2_{gp}}{2x2} + \frac{\sigma^2_{bgp}}{2x2} \quad (6)$$

olacaktır. Bu durumda G-katsayısı;

$$G - katsayısı = \frac{\sigma^2_b}{\sigma^2_b + \frac{\sigma^2_p}{2} + \frac{\sigma^2_{bp}}{2} + \frac{\sigma^2_{bg}}{2} + \frac{\sigma^2_{gp}}{2x2} + \frac{\sigma^2_{bgp}}{2x2}} \quad (7)$$

olarak elde edilir.

Diğer taraftan, eğer modelin mutlak olmasına karar verilmişse, o zaman mutlak hata

$$\sigma^2_{(\Delta)} = \frac{\sigma^2_p}{N_p} + \frac{\sigma^2_g}{N_g} + \frac{\sigma^2_{bp}}{N_p} + \frac{\sigma^2_{bg}}{N_g} + \frac{\sigma^2_{gp}}{N_p N_g} + \frac{\sigma^2_{bgp}}{N_p N_g} = \frac{\sigma^2_p}{2} + \frac{\sigma^2_g}{2} + \frac{\sigma^2_{bp}}{2} + \frac{\sigma^2_{bg}}{2} + \frac{\sigma^2_{gp}}{2x2} + \frac{\sigma^2_{bgp}}{2x2} \quad (8)$$

olacaktır.

Bu durumda güvenirlik (Phi)katsayısı;

$$\Phi - katsayısı = \frac{\sigma^2_b}{\sigma^2_b + \frac{\sigma^2_p}{2} + \frac{\sigma^2_g}{2} + \frac{\sigma^2_{bp}}{2} + \frac{\sigma^2_{bg}}{2} + \frac{\sigma^2_{gp}}{2x2} + \frac{\sigma^2_{bgp}}{2x2}} \quad (9)$$

olarak elde edilir.

Buradaki eşitliklerden de kolayca görülebileceği gibi, (7) eşitliğinin değeri (9) eşitiğinin değerinden daha büyüktür. Yine mutlak model için olan güvenirlik katsayısı daha küçük olmaktadır. İlk durum ile ikinci durum karşılaştırılırsa, göreceli modelde güvenirlik çaprazlanmış desende daha büyükken, mutlak modelde güvenirlik hem çaprazlanmış hem de yuvalanmış desen için aynı kalmaktadır.

3. D-Çalışması: Yukarıda yer alan her iki D-çalışmasında da tüm değişkenlik kaynakları tesadüfi olarak ele alınmıştı. Ancak bazı durumlarda bazı değişkenlik kaynakları sabit olarak düşünülebilir. Başka bir deyişle, araştırmacı çalışmasındaki örneklemin dışında bir genellemeye gitmek istemiyor olabilir. Bu durumda bir değişkenlik kaynağının sabitlenmesi, ölçme objesi (birey) ile bu değişkenlik kaynağı arasındaki etkileşime bağlı varyansın gerçek varyans olarak ele alınmasını gerektirir. Böylece sabitlenmiş bu değişkenlik kaynağının kendisine bağlı varyansı ise etkileşimsiz varyans bileşeni olarak ele alınır. Bu durum hatanın azalmasına ve böylelikle güvenirliğin artmasına sebep olur. Şimdiki D-çalışmasında puanlayıcı değişkenlik kaynağı sabitlenmiş olarak ele alınacaktır. Böylelikle öğrencilerin tüm olası puanlayıcılar tarafından puanlanmasıyla alacakları puanlarla ilgileniliyor olduğundan, bu durum bir bakıma gerçek dışıymış gibi görünebilir. Ancak bazı durumlarda puanlamayı yapan puanlayıcılar var olan ve mümkün olan en iyi puanlayıcılar olarak bilinirler ve bu puanlayıcıların verdikleri kararlar, itiraz edilemeyecek kadar doğru olarak düşünülebilir. Bu durumda da bu puanlayıcılar tüm evreni temsil eden uzmanlar olarak düşünülür ve daha ötesinde bir genellemeye gidilmek istenilmeyebilir.

Şimdi varsayılım ki yine iki puanlayıcı ve iki görevden oluşan tümüyle çaprazlanmış (b x g x p) bir desenimiz olsun. Buradaki tek fark "puanlayıcı" değişkenlik kaynağının sabitlenmiş olarak ele alınmasıdır.

Göreceli model kullanarak, göreceli hata varyansı;

$$\sigma^2_{(\delta)} = \frac{\sigma^2_{bg}}{2} + \frac{\sigma^2_{bgp}}{2x2} \qquad (10)$$

olarak elde edilir. G-katsayısı;

$$G - katsayısı = \frac{\sigma_b^2 + \frac{\sigma_{bp}^2}{2}}{\sigma_b^2 + \frac{\sigma_{bp}^2}{2} + \frac{\sigma_{bg}^2}{2} + \frac{\sigma_{bgp}^2}{2x2}} \quad (11)$$

şeklinde hesaplanır.

Eğer mutlak model kullanılırsa, mutlak hata varyansı:

$$\sigma^2_{(\Delta)} = \frac{\sigma_g^2}{2} + \frac{\sigma_{bg}^2}{2} + \frac{\sigma_{gp}^2}{2x2} + \frac{\sigma_{bgp}^2}{2x2} \quad (12)$$

olacaktır.Güvenirlik (Phi) katsayısı;

$$\Phi - katsayısı = \frac{\sigma_b^2 + \frac{\sigma_{bp}^2}{2}}{\sigma_b^2 + \frac{\sigma_{bp}^2}{2} + \frac{\sigma_g^2}{2} + \frac{\sigma_{bg}^2}{2} + \frac{\sigma_{gp}^2}{2x2} + \frac{\sigma_{bgp}^2}{2x2}} \quad (13)$$

şeklinde hesaplanır.

Yukarıdaki eşitliklerden kolayca görülebileceği gibi, 3. D-çalışmasındaki hem göreceli hem de mutlak modeldeki paylar ilk durumdakilerden daha büyüktür. Bu da göstermektedir ki; puanlayıcılar ile öğrenciler arasındaki etkileşime bağlı varyans bileşeninin gerçek varyans olarak düşünülmesi ve puanlayıcı varyansının ilişkisiz hata kaynağı olarak düşünülerek, hata varyansına ilişkin eşitliklerde yer almaması, güvenirlik katsayısının artmasına sebep olmaktadır. Böylece, bir değişkenlik kaynağının sabitlenmesinin katsayının artmasına neden olacağı söylenebilir.

Özetle, puanlayıcıların genellenebilirlik kuramı çerçevesinde (çaprazlanmış ya da yuvalanmış, göreceli ya da mutlak, tesadüfi ya da sabit) bir değişkenlik kaynağı olarak ele alınması, güvenirlik katsayısının kestirilmesini etkilemektedir. Ayrıca, tek değişkenlik kaynaklı ve tümüyle çaprazlanmış desenler için hesaplanan genellenebilirlik katsayısı, puanlayıcılar arası güvenirliğin tahmininde kullanılan (inter-class) sınıf içi korelasyon katsayısıyla aynı olmaktadır (Mushquash ve O'Connor, 2006).

1.8. MADDE TEPKİ KURAMI

Eğitimde ve psikolojide kullanılan ölçme araçlarının hazırlanmasında ve elde edilen puanların yorumlanmasında kullanılan klasik ölçme modelleri ve yöntemleri çok uzun yıllardır test uzmanlarına hizmet etmektedir. Ancak bu klasik model ve yöntemlerin ölçme araçlarında ve değerlendirmelerde kullanımı bazı sınırlılıkları da beraberinde getirmektedir. Bu sınırlılıklar; (a) ölçme aracındaki maddelerin özelliklerinin, ölçme aracının uygulandığı gruba ve (b) ölçmenin uygulandığı bireylerin yetenek seviyelerinin ölçme aracında bulunan maddelerin parametrelerine bağımlı olması şeklinde açıklanabilir (Hambleton, Swaminathan ve Rogers, 1991).

Psikometrisler bu tür genel ölçme uygulamalarında ortaya çıkan sorunların üstesinden gelebilecek yeni bir ölçme sistemini, madde tepki kuramı (MTK)'nı geliştirmişlerdir. 1980'lerde, madde tepki kuramı ölçme uzmanları tarafından çalışılan başlıca konulardan biri olarak karşımıza çıkmaktadır. Klasik test kuramı ve genellenebilirlik kuramı, gözlenen puanlara dayalı analizlerde bulunan yaklaşımlarken madde tepki kuramı bu tür bir yaklaşım değildir. Madde tepki kuramı, temelde bireyin gözlenemeyen yetenekleri ile ölçme aracındaki maddeler arasındaki ilişkiyi matematiksel bağıntılarla açıklamaya çalışan bir kuram olarak geliştirilmiştir. Madde tepki kuramında, yukarıda değinilen diğer modellerin sınırlılıkları ortadan kaldırılmaya çalışılmıştır.

Bireyin gözlenemeyen yetenek seviyesi, (ability level) yetenek seviyesinden bağımsız olduğu düşünülen uygulanan ölçme maddeleri ile gözlenebilir hale getirilmektedir. Bireyin gözlenen ve gözlenmeyen özellikleri arasındaki ilişki matematiksel olarak "normal ogive" ve "lojistik" fonksiyonlarla ifade edilir.

Madde tepki kuramı, iki temel varsayıma dayanır;

1) Bir ölçme aracı maddesindeki bir bireyin göstermiş olduğu performans, gizli özellikler ya da yetenekler denilen bir set faktörle tahmin edilir ya da açıklanır.
2) Bireylerin madde performansı ile madde performansı altında yatan özellikleri arasındaki ilişki madde karakteristik eğrisi (ICC) denilen monoton artan bir fonksiyon ile gösterilir.

Klasik test kuramından farklı olarak, madde tepki kuramı modelleri değişebilir modellerdir. Verilen bir madde tepki kuramı modelinin ilgilenilen ölçme aracı verisine uygunluğunu gösteren bazı özellikler bulunmaktadır: bireyin yetenek düzeyi ölçme aracı-bağımlı değildir ve madde özellikleri de grup-bağımlı değildir. Farklı madde setlerinden elde edilen yetenek tahminleri (ölçme hatası hariç) ve farklı gruplardan elde edilen madde parametreleri tahmini (ölçme hatası hariç) aynıdır. Madde tepki kuramında madde ve yetenek parametreleri değişmezdir. Bunlara ek olarak madde tepki kuramı, yetenek tahmininde her bir birey için ayrı bir standart hatanın kestirimine imkan tanır. Klasik test kuramında olduğu gibi tüm bireyler için tek bir standart hata tahmini vermez (Embretson ve Steven, 2000).

Madde tepki kuramında, bir maddenin doğru cevaplanma olasılığını veren lojistik dağılıma dayalı üç farklı model yer almaktadır. Bunlardan en basiti olan 1parametreli lojistik model aşağıdaki eşitllik ile gösterilir:

$$P(X_{is}=1|\theta_s,\beta_i)=\frac{e^{\alpha(\theta_s-\beta_i)}}{1+e^{\alpha(\theta_s-\beta_i)}}$$

Bu denklemde yer alan X_{is}=1, s bireyinin bir i maddesini doğru cevaplama olasılığını, $\theta_s-\beta_i$ ise s bireyinin yetenek düzeyi (θ_s) ile i maddesinin güçlük düzeyi (β_i) arasındaki farkı ifade eder. Bu denklemde yer alan α sabiti, madde ayırıcılık düzeyi için seçilebilecek her hangi bir sabit sayıyı temsil etmektedir. Bu sabit sayının 1 olarak seçildiği özel durumda, model aynı zamanda Rasch modeli olarak da tanımlanır (Embretson ve Steven, 2000). Madde tepki kuramında yer alan bir diğer model, 2 parametreli lojistik modeldir ve bu model aşağıdaki eşitlik ile ifade edilir:

$$P(X_{is}-1|\theta_s,\beta_i,\alpha_i)-\frac{e^{\alpha_i(\theta_s-\beta_i)}}{1+e^{\alpha_i(\theta_s-\beta_i)}}$$

Bu denklemin 1 parametreli lojistik model denkleminden tek farkı, eşitliğe α_i madde ayırıcılık parametresinin eklenmiş olmasıdır. 2 parametreli modele bir de şans başarısı (γ_i) parametreisinin eklenmesiyle 3 parametreli lojistik model denklemi elde edilir. Bu denklem aşağıda gösterildiği şekildedir:

$$P(X_{is} = 1|\theta_s, \beta_i, \alpha_i, \gamma_i) = \gamma_i + (1 - \gamma_i)\frac{e^{[\alpha_i(\theta_s - \beta_i)]}}{1 + e^{[\alpha_i(\theta_s - \beta_i)]}}$$

Genel olarak performansa dayalı ölçmelerde, madde tepki kuramında göz önünde bulundurulan farklı değişkenlik kaynakları vardır. Bu değişkenlik kaynaklarından biri *bireyler*dir. Birey değişkenlik kaynağı; bireylerin problem, görev ya da ürünle ilglili bilgi ya da becerisini ortaya koyduğu yetenekleridir ve değerlendirilen bireylerin yeteneklerinin birbirinden farklı olması beklenen bir durumdur. Performansa dayalı ölçmelerde ve yazılı yoklamalarda *puanlayıcılar* da diğer bir değişkenlik kaynağıdır. Performansın ölçülmesinde puanlayıcıların oldukça önemli bir yeri vardır. Çünkü puanlayıcıların puanlamalarını etkileyen farklı yaklaşım ve tepkileri bulunmaktadır. Bir başka değişkenlik kaynağı, bireyden yerine getirmesi beklenen *görev* ya da *madde*dir. Performanslara ilişkin görevler ya da maddeler, bireylerin performanslarını gözlenebilir hale getirmelerini sağlayacak kadar açık ve net, aynı zamanda bireyler arasındaki farklılıkları ortaya çıkarabilecek ayrıntıları içerecek nitelikte olmalıdır. Performansta yer alan görev ya da maddelerin sayısı ve parametreleri bireylerin performansını etkileyen özellikler arasındadır. Performansın ölçülmesinin gerekli olduğu durumlarda ölçmeyi oluşturan bu değişkenlik kaynaklarının ve birbirleriyle etkileşimlerinin incelenmesi gerekir. Bir bireyin ölçülebilir yeteneğinin, gerçekleştirmesi istenilen görevde gösterdiği performans ile kestirilebileceği varsayılır. Ancak bireylerin bu performanslarının gerçekte, puanlayıcıların farklı katılık düzeylerinde ya da maddeler ile görevlerin farklı güçlüklerde olmasıyla etkileşim içinde olduğunu da göz ardı etmemek gerekir.

Madde tepki kuramında, puanlayıcılar arası güvenirliğin ya da değişkenlik kaynağının birden fazla olduğu durumda güvenirlik hesaplanmasında kullanılabilen bir parametreli model, Rasch Modeli olarak bilinmektedir. Bir parametreli model, ilk olarak Danimarkalı bir matematikçi olan George Rasch tarafından geliştirilmiştir. Rasch modeli, yukarıda da ifade edildiği şekilde, sadece madde güçlük indeksi olan "β" parametresinin var olmasının yeterli görüldüğü, madde ayırıcılık indeksi olarak bilinen "α" parametrelerinin tüm maddeler için eşit ve bir kabul edildiği ve maddelerin şansla doğru yanıtlandığının göstergesi olan "γ" parametrelerinin tüm maddeler için sıfır olarak alındığı bir modeldir. Rasch modeli iki değişkenlik kaynaklı (facet) bir modeldir.

Bu modelde, değişkenlik kaynağı olarak genellikle bireyin örtük özelliği ile görevler ya da maddeler dikkate alınır.

1.9. ÇOK DEĞİŞKENLİK KAYNAKLI RASCH ÖLÇME MODELİ

Çok Değişkenlik Kaynaklı Rasch Ölçme Modeli (Multi-Faceted Rasch Measurement Model), Linacre (1989) tarafından geliştirilmiş, Rasch modelinin bir uzantısıdır. Orjinal Rasch modelinde bireylerin yetenekleri ve maddelerin güçlük düzeyleri aynı aralık ölçeği üzerinde yer alırken, çok değişkenlik kaynaklı Rasch ölçme modeli ikiden fazla değişkenlik kaynağını ele alarak Rasch modelinin daha da genişletilmiş bir halidir (Nakamura, 2000). Bu da, örneğin puanlayıcıların katı ya da cömertliği, farklı zamanlardaki ölçmelerin değişkenliği gibi farklı farklı değişkenlik kaynaklarının modelde yer almasına imkan tanır. Çok değişkenlik kaynaklı Rasch ölçme modelinde, puanlayıcıların katı ya da cömertliği, "puanlayıcıların, diğer puanlayıcıların verdikleri puanların ortalamasından daha yüksek ya da daha düşük yaptıkları sistematik puanlama" olarak ifade edilir (Engelhard ve Myford, 2003). Ayrıca bu durum *puanlayıcı etkisi* ya da *puanlayıcı hatası* olarak da tanımlanabilir.

Madde tepki kuramına dayalı ve bir parametreli lojistik modelin genişletilmiş hali olan çok değişkenlik kaynaklı Rasch ölçme modeli, bireylerin yetenek seviyelerini, görev ya da maddelerin güçlük düzeylerini, puanlayıcıların katılıklarını kestirmek için istatistiksel bir teknik sağlar. Çok değişkenlik kaynaklı Rasch ölçme modeli genel olarak üç değişkenlik kaynağı (bireyler, görevler ve puanlayıcılar) içerir ve aşağıda gösterilen şekilde ifade edilir:

$$log\left(\frac{P_{nijk}}{P_{nijk-1}}\right) = B_n - D_i - C_j - F_k$$

P_{nijk} : Bir bireyin (n) bir maddeye (i) gösterdiği performansa bir puanlayıcının (j), (k) kategorisinde verdiği puanın olasılığıdır.

P_{nijk-1}: Bir bireyin (n) bir maddeye (i) gösterdiği performansa bir puanlayıcının (j), (k-1) kategorisinde verdiği puanın olasılığıdır.

B_n: Bireyin (n) yetenek düzeyi.

D_i: Maddenin (i) güçlük düzeyi.

C_j: Puanlayıcının (j) katılık düzeyi.

F_k: k-1 kategorisinden k kategorisine geçişin güçlük düzeyi.

Seçilen örnekleme dayanan genellenebilirlik kuramından farklı olarak, çok değişkenlik kaynaklı Rasch ölçme modeli, her bir değişkenlik kaynağının kendi içinde olduğu kadar, değişkenlik kaynaklarının her birine ilişkin olarak örneklemden bağımsız parametre kestiriminde bulunmayı da sağlar. Değişkenlik kaynakları aynı zamanda analiz edilir ve bu analizler istatistiksel olarak birbirinden bağımsızdır. Bir başka deyişle, genellenebilirlik kuramı değişkenlik kaynakları arasındaki etkileşimi göz önünde bulunduran bir yaklaşımken, madde tepki kuramı etkileşimin olmadığı varsayımına dayalı bir yaklaşımdır. Değişkenlik kaynakları böylelikle genel bir doğrusal ölçek üzerinde, genellikle lojit ölçeğinde yer alırlar. Bireylerin yetenekleri, belirli maddelerin dağılımlarının özelliklerinden, belirli puanlayıcıların performansa verdikleri puanlardan vb. bağımsız olarak kestirilir. Benzer şekilde, maddelerin güçlük düzeylerinin ve puanlayıcılarının katılık/cömertlik düzeylerinin kestirimi de verilerin elde edildiği desende yer alan diğer değişkenlik kaynaklarının dağılım özelliklerinden bağımsız olarak yapılmaktadır (Smith ve Kulikowich, 2004). Bir diğer farklılık; genellenebilirlik kuramında ölçmenin kalitesi, bir grup ya da bir bütün olarak araştırılırken çok değişkenlik kaynaklı Rasch ölçme modeli, her bir değişkenlik kaynağının her bir elemanına ilişkin araştırma yapmaya imkan verir ve böylece her bir değişkenlik kaynağındaki elemanların (bireyler, maddeler ve puanlayıcılar) beklenmedik ya da aykırı durumlardaki performanslarının nasıl olduğunu belirleyebilmeyi sağlar (Alharby, 2006). Buna ek olarak çok değişkenlik kaynaklı Rasch ölçme modelinde her tür sistematik hata kaynağı bir değişkenlik kaynağı olarak görüldüğünden, genellenebilirlik kuramında değişkenlik kaynağı olarak ele alınmayan ölçme objeleri ya da bireyler de bir değişkenlik kaynağı olarak ele alınmaktadır (Alharby, 2006). Benzer şekilde, genellenebilirlik kuramında bir ölçme için bir güvenirlik katsayısı kestirilirken, çok değişkenlik kaynaklı Rasch ölçme modelinde her bir değişkenlik kaynağı için aynı anda ayrı ayrı güvenirlik katsayıları kestirilebilmektedir.

Çok değişkenlik kaynaklı Rasch ölçme modeli, parametrelerin ayrı ayrı kestirilmesine imkan tanır. Bireylerin yetenekleri, tüm puanlayıcıların tüm görevler ve/veya maddeler için verdikleri puanlardan kestirilir. Benzer şekilde herhangi bir maddenin güçlük düzeyinin kestirimi, tüm bireylerin bu maddedeki performanslarına ve tüm puanlayıcıların bu maddeye verdikleri puanlara, herhangi bir puanlayıcının katılık/cömertlik düzeyi de tüm maddelere ve tüm bireylere verdiği puanlara dayalı olarak gerçekleşir. Eğer veriler için değişmezlik (invariance) varsayımı sağlanmışsa bireylere ilişkin ölçmeler, puanlayıcılar ve maddeler arasındaki değişimden bağımsızdır. Böylece bireylere ilişkin ölçme sonuçları doğrusal ve objektif olacaktır. Eğer çok değişkenlik kaynaklı Rasch ölçme modeline dayalı analizlerin uygulanacağı veri küçük bir örneklemden elde edilmişse ya da boş bırakılmış (missing) veriler bulunmaktaysa, kestirimlerdeki kesinlik de o ölçüde azalacaktır.

Çok değişkenlik kaynaklı Rasch ölçme modeli analizinin FACETS bilgisayar programıyla iterasyon raporu, "facet haritası" ve her bir değişkenlik kaynağına ilişkin ayrı ayrı logit ölçme kestirimleri, standart hata ve model uyum istatistikleri gibi üç farklı istatistik sonuçları elde edilebilir (Linacre, 1989). Gerekli olan iterasyon sayısı, veriden iyi kestirimlerde bulunabilmek için verinin Rasch modeline ne kadar uyumlu olduğuna bağlıdır. Bir kaç iterasyondan sonra, kestirimler gittikçe düşmeye başlar. Hem normal yaklaşım algoritmi hem de koşulsuz maksimum olasılık algoritmi kullanılarak olabildiğince hızlı ve kesin kestirimlere ulaşılır (Linacre, 1994).

Çok değişkenlik kaynaklı Rasch ölçme modeli, tüm değişkenlik kaynaklarını içeren ve bu değişkenlik kaynaklarının birlikte araştırılmasına uygun olan bir desendir. Gözlenen puanların logoritmik olarak logit ölçeğine dönüşümüne dayalı doğrusal bir modeldir. Regresyon analizi terminolojisine göre; bağımlı değişken, başarılı kategoride olma olasılığı oranının logoritmik dönüşümü (log odds), bağımsız değişkenler ise değişkenlik kaynaklarıdır. Bazı araştırmacılar ise logit ölçeğini, ana etkiler (main effects) için bir log-doğrusal model olarak da tanımlarlar (Hetherman, 2004).

Verinin Rasch modeline uyum sağladığı kadarıyla, tüm değişkenlik kaynaklarına ilişkin ölçmeler tek bir cetvel ya da harita üzerinde ölçeklenebilmektedir. Cevaplama olasılıkları "log-odds" olarak ifade edilir ve ölçmelerin yer aldığı cetvel üzerindeki

ölçeğin birimi “log odds birimi” ya da “logits” olarak adlandırılır. Logits cetvelde yer alan yüksek pozitif değerlere sahip ölçmeler, bireyler için yüksek yetenek, puanlayıcılar için katı puanlama ve maddeler için ise yüksek güçlük düzeylerini ifade ederken, yüksek negatif değerler de düşük yetenekli bireyleri, cömert puanlama yapan puanlayıcıları ve düşük güçlük düzeyine sahip maddeleri göstermektedir. Bu logit ölçeğinin bir aralık ölçeği olması avantajının yanında, bir başka deyişle bir maddenin güçlük düzeyinin diğer bir maddedenin güçlük düzeyinden daha yüksek olduğunu göstermesinin yanı sıra, bir diğer avantajı da ne kadar daha güç olduğunu da doğrusal bir birim cinsinden verebiliyor olmasıdır.

Çok değişkenlik kaynaklı Rasch ölçme modelinde elde edilen bir diğer istatistik uyum istatistikleridir. Uyum istatistikleri, veri matrisindeki uyumsuzlukların derecesini gösterir, her bir madde için artıkların (residuals) büyüklüğüne ve yönüne (işaretine) ilişkin özetleyici bilgi sağlar. Artıklar, beklenen puanların gözlenen puanlardan çıkarılmasıyla hesaplanan değerlerdir. Veriler modelle uyumlu olduğunda, her bir artık değerin sıfır olması beklenir. Veri modele yaklaştığı ölçüde, standartlaşıtırılmış artığın kestirimi, ortalaması 0 ve varyansı yaklaşık 1 olacak şekilde az ya da çok normal dağılım gösterir. Böylece, verinin modele uyumu için kullanışlı bir ölçüt olduğundan, standartlaşıtırılmış artıkların yaklaşık normal dağılım gösterip göstermediğine bakılarak yorum yapılabilir (Hetherman, 2004).

Benzer şekilde dış uyum (OUTFIT) ve iç uyum (INFIT) istatistikleri, daha kolay yorumlanabilen uyum istatistikleri olarak incelenebilir. Her bir puanlayıcının puanlamadaki katılılığı/cömertliği, maddede yer alan diğer değişkenlik kaynaklarından bağımsız olarak kestirilir. Bu kestirimle ortaya çıkan yüksek değerler, puanlayıcının ortak ölçekte kendisinin altında kalan puanlayıcılardan daha katı olduğunun bir göstergesidir. Her bir puanlayıcının katılığının düzeyinin kestiriminden sonra, bu kestirimlerin verilere uygun olup olmadığı değerlendirilir. Benzer değlerlendirme her bir bireyin yeteneğinin kestirimi, her bir maddenin güçlük düzeyinin kestirimi ya da modeldeki diğer tüm değişkenlik kaynaklarının kestirimleri için yapılabilir (Hetherman, 2004).

Bir değişkenlik kaynağının modelle olan uyumu standartlaştırılmış artıkların karelerinin ortalaması olarak özetlenenir ve bu istatistik dış uyum (OUTFIT) istatistiği olarak tanımlanır. İç uyum (INFIT) istatistiğinin, dış uyum istatistiğinden tek farkı standartlaştırılmış artıkların kareler ortalmasının ağırlıklandırılmış hali olmasıdır. İç uyum ve dış uyum istatistiklerinden oluşan uyum istatistikleri her bir değişkenlik kaynağının her bir elemanı için hesaplanır. Puanlayıcı değişkenlik kaynağı için bu uyum istatistikleri, teorik modelle kestirilen puanlayıcı katılık parametrelerinin tutarlılığı için bir kanıt oluşturmaktadır. Dış uyum istatistikleri, aykırı (sapan) puanlara (outlier) karşı, iç uyum istatistiklerinden daha duyarlıdır. Bir başka deyişle, ağırlıklandırılmamış kareler ortalaması olan dış uyum istatistikleri, beklenmedik bir şekilde bir bireyin bir maddeyi çok kolay bulması ya da çok zor bulması durumlarına karşı çok hassastır. İç uyum istatistiği için istenilen değer 1'dir. 1'den büyük değerler puanlamada beklenenden daha fazla değişim (varyans) olduğunu, 1'den küçük değerler ise puanlamada beklenenden daha az değişim olduğunu, bir başka deyişle daha fazla bağımlılık olduğunu gösterir.

Alanyazında, hem iç uyum hem de dış uyum istatistiği için genel olarak kabul edilebilir değerlerin 0.5 ile 1.5 arasında olması önerilmektedir (Turner, 2003). Verinin çok değişkenlik kaynaklı Rasch ölçme modeline uyum gösterip göstermediğini belirlemek için iç uyum ve dış uyum istatistiklerinin [0.5, 1.5] aralığında olup olmadığı araştırılmalıdır. Değişkenlik kaynağı içindeki her bir elemana ilişkin iç uyum ya da dış uyum istatistik değerinin 1.5'ten büyük olması, anlamlı bir uyumsuzluk (misfit) olduğunun göstergesidir. Uyumsuz elemanlar, değişkenlik kaynağı içindeki tüm puan örtüntüsü içinde beklenmedik durum sergileyen elemanlardır. Örneğin; uyumsuz maddeler, kötü hazırlanmış olduklarının sinyalini verirler. Bu maddeler kendi içlerinde değerlendirildiklerinde belki de oldukça iyi hazırlanmış maddeler olabilirler, ancak madde örüntüsünde uyumsuzluk gösterirler. Uyumsuz bir puanlaycı, puanlayıcılar arasında göreceli olarak bir uyumsuzluk sergilemekte ve bu puanlayıcının puanlamadan çıkarılması ya da puanlayıcının tekrar eğitime alınmasına ilişkin bir işaret vermektedir. Uyumsuz bir birey ise, kendisinden beklenmedik şekilde bir maddeyi doğru ya da yanlış cevaplandırmıştır. Örneğin orta düzeyde başarı gösteren bir birey en zor iki maddeyi doğru ve en kolay maddeyi de yanlış cevaplandırmışsa, bu bireyin cevapları genel örüntüyle tutarlı değildir ve bu durum o bireyin uyumsuzluğuna işaret etmektedir.

Değişkenlik kaynağı içindeki her bir elemana ilişkin iç uyum ya da dış uyum istatistik değerinin 0.5'ten küçük olması, anlamlı olarak uyum üstü bir durumun söz konusu olduğuna işarettir. Uyum üstü elemanlar, değişkenlik kaynağının genel örüntüsü içinde bu elemanların değişkenlik göstermediğinin ya da değişkenlik kaynağındaki diğer elemanlardan bağımsız olarak bir bilgi sağlamadıklarının göstergesidirler. Örneğin, uyum üstü maddeler, sadece en üstteki öğrencilerin doğru cevaplandırdığı maddeler olarak gözlenebilirler (Hetherman, 2004).

Çok değişkenlik kaynaklı Rasch ölçme modelinin, verinin modele ne kadar uyumlu olduğunu belirlemeyi sağlayan bir diğer istatistiği de ayırma indeksi (G)'dir. Ayırma indeksi, değişkenlik kaynağındaki objelerin birbirinden ne kadar ayrıldığını gösteren bir ölçüdür. Bireyler ve maddeler için bu ölçünün büyük, puanlayıcılar için ise sıfıra yakın bir değer çıkması beklenir. Yüksek değere sahip bir ayırma indeksi, puanlayıcıların bir birlerinden ne kadar farklı puanlama yaptıklarının bir göstergesidir. Puanlayıcı değişkenlik kaynağı için elde edilen düşük ayırma indeks değerleri, istenen bir durum olarak, katılık/cömertlik açısından küçük farklılıklar olduğunun bir göstergesidir. Ayırma indeksi (G) ayrıca, her bir değişkenlik kaynağı için, değişkenlik kaynağında yer alan elemanların kaç farklı düzeye ayrılabildiğini gösteren (4G+1)/3 formülüyle hesaplamada da kullanılmaktadır (Lee ve Kantor, 2003; Hetherman, 2004).

Parametre kestirimlerinin ne kadar iyi yapıldığını yorumlamada fayda sağlayacak bir başka istatistik ayırma indeksi güvenirliğidir. Klasik güvenirlik katsayısında olduğu gibi ayırma indeksi güveirliği de 0 ile 1 arasında değişen değerler almaktadır. Bu değer, değişkenlik boyutundaki elemanların bir birlerinden ne kadar güvenilir bir şekilde ayrıldıklarını gösterir. İlgilenilen değişkenlik kaynağı (birey ya da öğrenci değişkenlik kaynağı olup genellenebilirlik kuramında bu değişkenlik kaynağına "ölçme objesi" denmekteydi) için elde edilen ayırma indeksi güvenirliği; klasik kuramdaki KR-20, Cronbach alfa veya genellenebilirlik katsayısına eş değerdir ve iç tutarlılık anlamıda yorumlanabilir (Nakamura, 2000). Birey, madde ve görev gibi değişkenlik kaynakları için ayırma indeksi güvenirliğinin bire yakın bir değer çıkması istenir. Puanlayıcı değişkenlik kaynağı için ise ayırma indeksi güvenirliğinin bire yakın çıkması, puanlayıcıların katılık/cömertlik açısından birbirlerinden ne kadar güvenilebilir şekilde farklı olduklarını gösterir. Puanlayıcı değişkenlik kaynağı için bu istenen bir durum

değildir. Linacre (1989)'ye göre, puanlayıcı değişkenlik kaynağı için ayırma indeksi güvenirliği, puanlayıcılar arası uyum (agreement) indeksinin hesaplanmasında "1 - ayırma indeksi güvenirliği" şeklinde kullanılır. Bu durumda ayırma indeksi güvenirliği ne kadar düşük çıkarsa, puanlayıcılar arası uyum o kadar büyük olacaktır.

FACETS programında yorumlanan bir diğer istatistik X^2 (Ki-kare)'dir. X^2 istatistiği her bir değişkenlik kaynağındaki elemanların birbirlerinden farklılıklarının istatistiksel olarak anlamlı olup olmadığının test edilmesinde kullanılır. Bu testte kurulan yokluk hipotezi örneğin; "çalışmada yer alan puanlayıcıların katılık/cömertlikleri birbirine eşittir" şeklindedir. X^2 testinin manidar çıkması elemanlar arasında anlamlı farklılıkların olduğunu gösterir.

Bu çalışmada uygulanan matematik performansının ölçülmesinde kullanılan maddeler, 1999 yılında ülkemizde uygulanmış olan TIMSS çalışmasından alınmıştır. TIMSS çalışması hakkında açıklayıcı bilgi aşağıda verilmeye çalışılmıştır.

TIMSS ilk olarak 3. Uluslararası Matematik ve Bilim Çalışması (3. International Mathematics and Science Study) olarak isimlendirilmesine karşın son yıllarda Ululararası Matematik ve Bilim Çalışması Eğilimi (The Trends in International Mathematics and Science Study) adı altında uygulanan standart bir ölçme aracı niteliği taşıyan, uluslar arası kapsamda bir ölçme çalışmasıdır (TIMSS, 2008). Eğitim politikasını belirleyen, öğretim programını hazırlayan uzmanlara ve araştırmacılara, eğitim sistemini değerlendirebilmeleri açısından bilgi sağlamak amacıyla uygulanmış olan TIMSS, ilk kez 41 ülkede beşinci sınıf öğrencilerinin matematik ve fen bilgisi başarılarını karşılaştırmıştır. İlk TIMSS uygulaması 1994 - 1995 yıllarında gerçekleştirilmiştir. TIMSS sonuçları 1996 yılında ilk kez rapor edilmiş, sonuçlar bir tartışma ortamı yaratmış, reform çalışmalarını hızlandırmış ve dünya çapında eğitim uzmanlarına, araştırmacılara ve karar vericilere değerli bilgiler sağlamıştır (TIMSS 1999 Türkiye Raporu, 2003, s.1).

TIMSS-1999, 1998 - 1999 yılında uluslararası düzeyde sekizinci sınıf öğrencilerinin matematik ve fen bilgisi başarılarını, TIMSS 1995 uygulamasına göre gelişimini izleyebilmek üzere tasarlanmıştır. Ayrıca, ilk TIMSS uygulamasından itibaren geçen dört yıllık süreçteki öğrenci gelişimini, ilk uygulamada dördüncü sınıfta olup sekizinci

sınıfa gelmiş olan öğrenciler üzerinden görebilmeyi hedeflemiştir. Böylece TIMSS-1999 geçen süreç içinde öğrenci başarısındaki değişim hakkında da bilgi sağlayabilmiştir (TIMSS, 2003, s.1).

TIMSS çalışmasında 1999 yılında katılan ülkemiz ve diğer yeni katılan ülkeler, hem öğrencilerin gelişimini izleyebilme hem de diğer ülkelere göre başarı düzeylerini karşılaştırmalı olarak değerlendirebilme şansına sahip olmuştur. Ancak ülkemiz 2003 yılında yapılan çalışmaya katılmama kararı almıştır. Bu nedenle ülkemizin 1999'dan 2003'e kadar olan süreçteki gelişimini görme imkanı olmamıştır. Ülkemizin de katıldığı 2007'de uygulanmış olan TIMSS'in raporlaştırılma çalışması halen devam etmektedir.

TIMSS başarı testleri genel olarak okullarda izlenen öğretim programında ele alınan temel becerilere odaklanmaktadır. Matematik için hazırlanan testin kapsamında; kesirler ve sayıları anlama, ölçme, veri gösterimi-analiz ve olasılık, geometri ve cebir konuları yer almaktadır. TIMSS-1999 uygulamasında matematik başarısına göre, ülkemiz projeye katılan 38 ülke arasında 31. sırada, ilk sırada Singapur son sırada ise Güney Afrika yer almıştır (TIMSS, 2003, s.3-4).

TIMSS'de yer alan yaklaşık 300 maddenin dörtte biri, öğrencilerin cevapları kendilerinin vermesini gerektiren açık uçlu sorulardan oluşmaktadır. Bu cevapların puanlanmasında iki rakamlı tanımlayıcı kodlamadan oluşan bir dereceli puanlama anahtarı (rubrik) kullanılmaktadır. Bu kodlamadaki ilk rakam cevabın doğruluğunu ikinci rakam ise öğrencinin genel yaklaşımını ya da yanlışlığını gösterecek şekilde cevaplara ilişkin belirli bir durumu ifade eder. TIMSS'de yer alan açık uçlu soruların çoğunluğu kısa cevaplı maddelerden oluşmaktadır ve her bir maddenin değeri 1 puan değerindedir. Daha uzun cevap gerektiren diğer açık uçlu maddeler ise 2 ya da 3 puan değerindedir. Açık uçlu cevapların puanlanmasında kullanılan dereceli puanlama anahtarı İngilizce olarak geliştirilmekte ve her katılımcı ülkenin ana diline çevrilmektedir. Tek tip kodlamanın oluşturulabilmesi TIMSS'in kalite kontrolü için en önemli konulardan biridir. Öğrenci örnekleri ile hazırlanan detaylı kodlama kılavuzlarının yanı sıra her bir ülkenin kodlamasına yardımcı olması amacıyla uluslar arası bilgilendirme seminerleri de verilmektedir. TIMSS'in puanlamadaki kalitesinden emin olmak amacıyla hem ülkeler için hem de ülkeler arası açık uçlu soruların

kodlanmasının puanlayıcılar arası güvenirliğinin belirlenmesi çalışmaları yapılmaktadır (Teresa, 1997).

TIMSS, gelişim politikası kapsamında kamu sorumluluğunun oluşturulmasında, başarının belirlenmesinde ve tanımlanmasındaki ilerleme ya da gerileme alanlarının saptanmasında ve adil eğitim sistemi konularında bilgi sağlamaktadır. Dünyanın her yerinden yaklaşık 50 ülke TIMSS'e katılmaktadır. IEA (International Association for the Evaluation of Educational Achievement) projesinin Armsterdam'da yer alan merkezi, katılımcı ülkelerden temsilcilerin ve organizasyonların dünya çapındaki ortaklığından oluşan Boston Kolej'deki TIMSS Uluslararası Öğrenci Merkezi'nce yürütülmektedir (TIMSS, 2008).

TIMSS dışında ülkemiz uluslararası alanda ayrıca iki önemli uygulamaya daha katılmaktadır. Bunlardan biri PISA projesidir. Ülkemizin ilk dönemde (1997-2000) yer almayıp, 2000-2003 döneminde katıldığı PISA, Uluslararası Öğrenci Değerlendirme Projesi (Program for International Assessment), 15 yaş grubu öğrencilerinin ağırlıklı olarak matematik başarısına yönelik hazırlanmış olmakla birlikte, öğrencilerin fen bilimleri, okuma ve problem çözme becerilerinin de ölçülmesini de kapsamaktadır. Projede Türkiye dahil 41 ülke yer almıştır. Diğer bir uluslararası uygulama ise PIRLS projesidir. Uluslararası Okuma Becerilerinde Gelişim Projesi (Progress in International Reading Literacy Study) ilk kez 2001 yılında, ülkemizin de aralarında bulunduğu 35 ülkede uygulanmıştır. Beş yıllık dönemlerle tekrarlanmakta olan PIRLS'nin amacı ilköğretim 4. sınıf (9 yaş) öğrencilerinin okuma becerilerinin seviyesini ve zaman içindeki gelişimlerini gözlemleyebilmektir. 2001 yılında uygulanan PIRLS sonucunda ülkemiz 35 ülke arasında 28. sırada yer almıştır (PISA 2003 Projesi, Ulusal Nihai Raporu, 2005).

1.10. PROBLEM DURUMU

Ölçme, eğitim alanında uygulanan programların başlangıcında, program sürecinde ve sonucunda çok önemli bir yere sahiptir. Eğitim programı sürecinin farklı periyotlarında uygulanan ölçme durumları farklı biçimlerde karşımıza çıkmaktadır. Amaca yönelik olarak zaman zaman çoktan seçmeli, kısa cevaplı ya da tamamlamalı, doğru-yanlış gibi ölçme araçları, zaman zaman ise daha karmaşık durumlar için performansın ölçülmesi

olarak karşımıza çıkmaktadır. Genellikle hem kapsam geçerliliğini daha iyi sağlaması hem puanlanmasının objektif oluşu hem de puanlamanın kolay ve az zaman alır olması nedeniyle çoktan seçmeli testler daha yaygın bir kullanıma sahiptir. Ancak becerilerin sergilenmesini gerektiren öğrenim ürünlerinin ve çoktan seçmeli testler ile ölçülemeyen diğer performansların ölçülmesine daha olası bir yöntem olarak görünmesi, öğretim programı ile öğretme arasındaki bağı kurmakta daha etkin olması ve gerçek yaşam becerileri ile daha yakından ilişkili olması nedeniyle performansın ölçülmesi, 1990'larda çarpıcı bir biçimde tekrar görünmeye başlamıştır. Fakat "kişilerin bireysel kararları üzerine kurulu olan performansın ölçülmesi; güvenirliği, geçerliği sağlayabilmekte midir?" sorusu da performansın ölçülmesinde önemli bir problem olarak araştırmacıların karşısına çıkmaktadır.

Performansın ölçülmesi, psikometri ile uğraşan bilim adamlarını, test geliştiricileri ve araştırmacıları "ne tür bir sonuç elde etmek istiyoruz?"a ilişkin kanıtlar ortaya çıkaracak gözlenebilir durumlar tasarlanması, bu kanıtların özetlenmesinin nasıl öğrenileceği ve ölçme sistemlerinin nasıl geliştirilebileceği konusunda zorlamaktadır (Brown, W. L., O'Gorman, K. ve Du, Y., 1996). İşte bu tür bir kısıtlılığa çözüm olması amacıyla da ölçme sürecine birden fazla puanlayıcının katılması tercih edilebilir bir çözüm olabilmektedir. Ancak bu kez de güvenirlik çalışmaları için birden çok değişkenlik kaynağının birlikte dikkate alınması gerekliliği ortaya çıkar. Tüm değişkenlik kaynaklarının birlikte göz önüne alındığı kapsamlı bir güvenirlik katsayısının elde edilmesinde klasik test kuramına dayalı ölçme yaklaşımlarının yanısıra genellenebilirlik kuramı ve çok değişkenlik kaynaklı Rasch ölçme modeli de daha yeni yaklaşımlar olarak yer almaktadırlar. Bu iki yeni yaklaşım, özellikle son yirmi yılda araştırmalarda ve uygulamalarda kullanılan yöntemler olarak karşımıza çıkmakla birlikte sıklıkla kullanılmamaktadırlar. Her iki yaklaşımın kullanıldığı araştırmalara bakıldığında genellikle hipotetik verilerin kullanıldığı sınırlı durumlar karşımıza çıkmaktadır.

Genellenebilirlik kuramı ve değişkenlik kaynaklı Rasch ölçme modeli ile gerçek veriler kullanılarak yapılan çok az sayıda araştırma bulunmakta ve bu araştırmaların çoğunluğu da sadece bu yaklaşımlardan birinin kullanıldığı uygulamaları içermektedir. Ayrıca klasik test kuramına dayalı ölçme yaklaşımlarının, genellenebilirlik kuramının ve çok değişkenlik kaynaklı Rasch ölçme modelinin birlikte kullanıldığı ülkemizde yapılan bir

araştırmaya rastlanmamıştır. Bu açıdan bakıldığında tüm bu yaklaşımların birlikte yer aldığı ve gerçek verilerin kullanıldığı bir çalışmanın yapılmasına gereksinim olduğu söylenebilir.

Performansın ölçülmesinde kullanılan maddelerin belirli bir dereceli puanlama anahtarına dayalı olarak birden fazla puanlayıcı tarafından puanlandığı ve elde edilen puanların da yukarıda belirtilen tüm yaklaşımlar dahilinde gerçek veriler üzerinden değerlendirilmesi, üç kuramın sınırlı ve avantajlı yönlerinin kıyaslanması açısından bu çalışmanın önemli olduğu düşünülmektedir.

Bu araştırmada, performansın ölçülmesinde kullanılan maddelerin bütünsel bir dereceli puanlama anahtarı ile birden fazla puanlayıcının puanlaması sonucu elde edilen puanlara klasik test kuramına dayalı ölçme yaklaşımı, genellenebilirlik kuramı ve çok değişkenlik kaynaklı Rasch ölçme modelinin birlikte uygulanması ile hangisinin(lerinin) daha kullanılabilir olduğu, avantajlı olanının ve daha çok bilgi sağlayanının belirlenmesi amacıyla aşağıdaki problem cümlesi ve alt problemlere cevap aranmaktadır.

1.11. PROBLEM CÜMLESİ

Ölçme sürecinde; performansın ölçülmesinde kullanılan maddelerin bütünsel bir dereceli puanlama anahtarı ile birden fazla puanlayıcı tarafından puanlanması sonucu elde edilen puanların klasik test kuramına dayalı ölçme yaklaşımı, genellenebilirlik kuramı ve değişkenlik kaynaklı Rasch ölçme modeli ile kestirilen parametreleri nasıldır?

1.12. ALT PROBLEMLER

I. Klasik test kuramına göre;

    1. Matematik performansının ölçülmesinde yer alan maddeler, öğrencilerin matematik başarısını var olan boyutta ölçmekte midir?
    2. Matematik performansının ölçülmesinin iç tutarlılık düzeyi nedir?
    3. Matematik performansının ölçülmesinde dört farklı puanlayıcı arasındaki tutarlılık düzeyi nedir?

II. Genellenebilirlik kuramına göre;

1. Matematik performansının ölçülmesinde kestirilen genellenebilirlik kuramı parametreleri nelerdir?

1.a. Tek değişkenli (performansın ölçülmesi tek boyuttan oluştuğu için) genellenebilirlik (G) çalışması sonucunda matematik performansının ölçülmesiyle kestirilen varyans ve toplam varyansları açıklama yüzdeleri nelerdir?

1.b. Orjinal puanlayıcı ve madde sayıları ile puanlayıcı sayılarının bir arttırılıp bir azaltılması ve madde sayılarının üçte bir arttırılıp azaltılması senaryoları sonucunda karar (K) çalışmasıyla kestirilen,

1.b.a. G katsayıları nasıl değişmektedir?

1.b.b. Phi katsayıları nasıl değişmektedir?

III. Çok değişkenlik kaynaklı Rasch ölçme modeline göre;

1. Model veri uygunluğu nasıldır?
2. Bireylerin yetenekleri ve uygunluk istatistikleri nasıldır?
3. Puanlayıcıların katılık/cömertlik istatistikleri nasıldır?
4. Madde güçlükleri ve uygunluk istatistikleri nasıldır?

IV. Klasik test kuramı, genellenebilirlik kuramı ve çok değişkenlik kaynaklı Rasch ölçme modeline göre birey (öğrenci) ve puanlayıcı değişkenlik boyutlarında elde edilen parametrelerin karşılaştırılması nasıldır?

1.13. ARAŞTIRMANIN AMACI

Bu araştırmada performansın ölçülmesinde yer alan maddelerin birden fazla sayıdaki puanlayıcı tarafından bütünsel bir dereceli puanlama anahtarı ile puanlanması halinde birden çok değişkenlik kaynağını birlikte yer aldığı ölçme durumunda, klasik test kuramına dayalı ölçme yöntemi, genellenebilirlik kuramı ve çok değişkenlik kaynaklı

Rasch ölçme modeli ile gerçek veriler kullanılarak kestirilen parametreler araştırılarak bu üç yaklaşım kıyaslanmaktadır.

Aynı veriler kullanılarak klasik test kuramına dayalı ölçme yöntemi, genellenebilirlik kuramı ve çok değişkenlik kaynaklı Rasch ölçme modelinden kestirilen parametrelerin karşılaştırılarak gerçek verilerle hangi yaklaşımın avantajlı olduğu, daha çok bilgi sağladığı ve iyi işleyip işlemediği ile bu yaklaşımların ürettiği bilgilerin birbirleriyle tutarlılıkları araştırılmaktadır. Bu doğrultuda araştırmada 2007 yılında uygulanan ve değişkenlik kaynağı olarak birey, puanlayıcı ve maddelerin bulunduğu gerçek bir ölçme uygulamasından elde edilen verilerle klasik test kuramına dayalı ölçme yöntemi, genellenebilirlik kuramı ve çok değişkenlik kaynaklı Rasch ölçme modeli analizleri yapılarak elde edilen parametreler incelenmektedir. Bu amaçla;

1. Klasik test kuramına dayalı ölçme yöntemiyle elde edilen parametreleri belirleyerek gerçek verilerle kullanabilirliğini saptamak,

2. Genellenebilirlik kuramı ile elde edilen parametreleri belirleyerek gerçek verilerle kullanabilirliğini belirlemek,

3. Çok değişkenlik kaynaklı Rasch ölçme modeli ile elde edilen parametreleri belirleyerek gerçek verilerle kullanabilirliğini saptamak,

4. Performansın ölçülmesine ilişkin yapılan uygulamanın, bu üç yaklaşım ile birey ve puanlayıcı değişkenlik boyutlarında elde edilen parametrelerinin karşılaştırılması araştırmanın içeriğinde yer almıştır.

1.14. ARAŞTIRMANIN ÖNEMİ

Bu araştırmayla, benzer ölçme durumları için öne sürülmüş olan klasik test kuramına dayalı ölçme yöntemi, genellenebilirlik kuramı ve çok değişkenlik kaynaklı Rasch ölçme modeli yaklaşımları kullanılarak birbirleriyle ve kendi içlerinde tutarlılıkları sınanarak kuramsal bir katkı sağlanabileceği düşünülmektedir. Böylelikle bu üç yaklaşımın kuramsal yapısının gerçek yaşam durumlarında aynı verilerle test edilmesi yoluyla uygulanabilirliklerinin, üstünlük ve sınırlılıklarının belirlenmesi ve

tutarlılıklarının ortaya çıkarılması ile kuramların gelişimine ve tartışmalara katkı sağlaması umulmaktadır.

Ayrıca üç yaklaşımdan hangisinin kullanılmasının daha çok bilgi sağladığı ve kullanılabilir olduğunun belirlenmesi ile uygulamacıların pratikte tercih edebilecekleri yaklaşımı seçmeleri konusunda bu araştırmanın bilimsel bir kaynak olabilmesi beklenmektedir.

Matematik eğitimcileri, matematikle ilgili çoktan seçmeli sınavlar yerine performansın ölçülmesi gibi daha subjektif ölçme araçları kullandıklarında, öğrencilerinin matematiksel bilgileri hakkında karar verirken kullandıkları puanların ne kadar güvenilir olduğunu bilmek isterler. Bunun yanısıra matematik performansının ölçülmesinde en etkili olan değişkenlik kaynağının ne olduğunu ve ölçme hatasını en aza indirmek için ölçmenin nasıl yapılması gerektiğini de bilmeye ihtiyaç duyarlar. Aynı zamanda bu noktalara temas edilerek, matematik eğitimcilerini bu konularda aydınlatmak çalışmanın bir diğer amacı olmuştur.

1.15. SAYILTILAR

Araştırmada kullanılan 1999-TIMSS'de yer almış olan 24 maddenin altı farklı ilköğretim okulu 8. ve 9. sınf öğrencilerine uygulanması sonucu elde edilen verileri, puanlayıcılar birbirlerinden bağımsız olarak puanlamışlardır.

1.16. SINIRLILIKLAR

2007 yılında uygulanan 18 açık uçlu TIMSS soruları değişkenlik kaynaklı Rasch ölçme modeli ile 203 öğrenci, 18 madde ve 4 puanlayıcı için birey x madde x puanlayıcı karşılıklı etkisi olası yanlılık kombinasyonu (203x18x4) 14.616'dır. Bu araştırmada çok değişkenlik kaynaklı Rasch ölçme modeli analizleri için sadece yukarıda verilen sayıda puanlayıcı, madde ve öğrenci kullanılmıştır ve çalışma yalnızca buradan elde edilen veriler üzerindeki birey x madde x puanlayıcı karşılıklı etkisi ile sınırlı kalmıştır.

1.17. KISALTMALAR

KTK: Klasik Test Kuramı
MTK: Madde Tepki Kuramı
G Kuramı: Genellenebilirlik Kuramı
G Çalışması: Genellenebilirlik Çalışması
K Çalışması: Karar Çalışması
G Katsayısı: Genellenebilirlik Katsayısı
Phi Katsayısı: Güvenirlik Katsayısı
ÇDKRÖM: Çok Değişkenlik Kaynaklı Rasch Ölçme Modeli
b: birey (öğrenci)
p: puanlayıcı
m: madde
b x p: birey - puanlayıcı ortak etkisi
b x m: birey - madde ortak etkisi
p x m: puanlayıcı - madde ortak etkisi
b x p x m: birey – puanlayıcı – madde ortak etkisi

1.18. İLGİLİ ARAŞTIRMALAR

Bu bölümde, araştırmanın kapsamını oluşturan performansın ölçülmesinde, birden fazla değişkenlik kaynağının birlikte değerlendirilmesine olanak sağlayan genellenebilirlik kuramı ve çok değişkenlik kaynaklı Rasch ölçme modelinin kullanıldığı çalışmalar, yurtiçi ve yurtdışı araştırmalar olmak üzere iki başlık altında özetlenmiştir.

1.18.1. Yurt İçinde Yapılan Araştırmalar

Yelboğa (2007), klasik test ve genellenebilirlik kuramına göre güvenirliğin bir iş performansı ölçeği üzerinde incelenmesi adlı çalışmasında, 2005 ve 2006 yıllarında, iş performansı ölçeği kullanarak elde ettiği gerçek verilerin klasik test kuramı ve genellenebilirlik kuramına göre güvenirliklerini kestirmiş ve puanlayıcılar arası tutarlılıkları belirlemeye çalışmıştır. Klasik test kuramında; test tekrar-test ve Cronbach alfa güvenirlik katsayılarını, puanlayıcılar arası güvenirlik için Kendall'ın uyum katsayısı ve genellenebilirlik kuramında ise çok değişkenlik kaynaklı modelle G ve Phi

katsayılarını hesaplamıştır. Araştırmadan elde edilen sonuçlara göre, klasik test kuramında ve genellenebilirlik kuramının çok değişkenlik kaynaklı modeliyle elde edilen güvenirlik katsayılarının birbiriyle uyumlu sonuçlar verdiği görülmüştür. İş performansı ölçeğiyle birden fazla puanlayıcıyla puanlama yapılan durumlarda genellenebilirlik kuramının kullanılmasının uygun olacağı belirtilmiştir.

Atılgan (2004), genellenebilirlik kuramı ve çok değişkenlik kaynaklı Rasch modelinin karşılaştırılmasına ilişkin bir araştırma başlıklı çalışmasında, genellenebilirlik kuramı (G Kuramı) ve çok değişkenlik daynaklı Rasch modeli (ÇDKRÖM)'nin karşılaştırılmasına ilişkin bir araştırma yapmıştır. Betimsel bir araştırma niteliği taşıyan çalışma, 2002 ve 2003 yıllarında yapılan müzik öğretmenliği özel yetenek seçme sınavının birinci aşamasına ilişkin 7 boyut ve bu boyutlara ilişkin 28 görev üzerinden elde edilen veriler kullanılarak yapılmıştır. Araştırmanın sonucunda, genellenebilirlik kuramı ve çok değişkenlik daynaklı Rasch modeli yaklaşımları ile elde edilen sonuçların her durumda paralellik göstermediği, durumdan duruma farklılaşabildiği gözlenmiştir. Ancak özellikle puanlayıcı ve görev değişkenlik kaynağı için her iki yaklaşımın daha tutarlı sonuçlar ürettiği de belirtilmiştir. Ayrıca, aynı ölçme durumunun alt boyutlarından oluşan testlerde, genellenebilirlik kuramının tek ve çok değişkenlik kaynaklı modellerinin farklı sonuçlar ürettiği, genellenebilirlik (G) ve güvenirlik (Phi) katsayılarının modelden etkilendiği ve bu katsayıların alternatif karar çalışmalarıyla elde edilmesi ile gerçek durumda kestirilmesi arasında farklılık olduğu gözlenmiştir.

Atılgan (2005), genellenebilirlik kuramı ve puanlayıcılar arası güvenirlik için örnek bir uygulama verdiği çalışmasında, genellenebilirlik kuramını tanımlamış, kuramla ilgili temel kavramları açıklamış, klasik test kuramına göre üstün olan yönlerini vurgulamış ve uygulanmasını da hipotetik olarak örneklendirmiştir. Yaptığı bu çalışmada, genellenebilirlik kuramının güçlü bir istatistiksel temeli olduğunu ve varyans analizine dayandığını vurgulamıştır. Puanlayıcılar arası güvenirliğin belirlenmesi gibi hata kaynağının birden çok olduğu ölçme durumlarında, ölçmenin psikometrik özelliklerinin belirlenmesinde, ölçme araçlarının geliştirilmesinde, en uygun madde ve puanlayıcı sayısının belirlenmesinde genellenebilirlik kuramının klasik test kuramının yerine kullanılmasının daha uygun olabileceği sonucuna varmıştır.

Yurt içinde yapılan sınırlı sayıda araştırmalara ilişkin olarak, performansın ölçülmesinde güvenirliğin belirlenmesinde genellenebilirlik kuramı tercih edilebilir bir yaklaşım olarak görülmektedir. Performansın ölçülmesinde, hem birden fazla değişkenlik kaynağının birlikte göz önüne alınarak güvenirliğin belirlenmesi hem de karar çalışmaları ile farklı sayıda değişkenlik kaynaklarına ilişkin önceden bilgi vermesi açısından genellenebilirlik kuramının kullanılması önerilmektedir.

1.18.2. Yurt Dışında Yapılan Araştırmalar

1.18.2.1. Genellenebilirlik Kuramına Dayalı Araştırmalar

Lei, Smith ve Suen (2007) tek bir deneği gözlemleme çalışmasından elde edilen verinin güvenirliğinin kestiriminde genellenebilirlik kuramını kullanmanın sınırlılıklarını araştırmışlardır. Klinik ve eğitim durumlarında kullanılan ölçme yöntemlerinden birinin "davranışları doğrudan gözlemlemek" olduğu belirtilmiştir.Yaptıkları çalışmada, gözleme dayalı ölçmelerde verilerin, tekrarlı ölçmeler yapılarak ve küçük örneklemler kullanılarak elde edilidiği vurgulanmış, bu verilerin güvenirliğini kestirimenin zor olduğu ifade edilmiştir. Gözleme dayalı araştırma desenleri için görev ya da madde, puanlayıcı, ölçme durumu gibi değişkenlik kaynakları ve ölçme objesine (genellikle bireylerdir) dayalı güvenirlik tanımları açıklanmıştır. Bu tür çalışmaların güvenirliği, genellenebilirlik kuramının kullanımına ilişkin teorik ve pratik problemler tartışılmıştır. Tek bir deneğin yer aldığı gözleme dayalı çalışmalarda, genellenebilirlik kuramının değişkenlik kaynaklarının çok iyi açıklandığı şartlar altında kullanılabileceği vurgulanmıştır. Bu düşünceyi açıklanmak üzere iki sayısal örneğe yer verilmiştir. İlk örnekte tek bir denek üzerinde yapılan ölçmede, ölçme objesinin birey değil ölçme durumları (occasions), değişkenlik kaynağının gözlem yapan puanlayıcılar olması gerektiği vurgulanarak, 64 farklı ölçme durumu ve bir puanlayıcıya ilişkin G ve D çalışmaları yapılmıştır. Elde edilen sonuçlara göre bir gözlemci yerine iki gözlemci kullanılarak göreceli değerlendirmeye ilişkin G katsayısının 0.68'den 0.81'e ve mutlak değerlendirmeye ilişkin Phi katsayısının 0.66'dan 0.79'a arttığı bulunmuştur. Çalışmada yer alan ikinci örnekte, tek bir denek yerine çok küçük bir örneklemin seçildiği durumlarda da ölçme objesinin ölçme durumları olması gerektiği vurgulanmış, dört birey ve on farklı ölçme durumu üzerinde G ve D çalışmaları yapılmıştır. Bu örnekte, ilk örnekte elde edilen sonuçlara paralel olarak, puanlayıcı sayısının birden ikiye

arttırılması durumunda genellenebilirlik ve güvenirlik katsayılarında artış gözlendiği vurgulanmıştır. Çalışmanın sonucunda, tek denekli gözlemlerde ölçme objesinin birey değil ölçme durumları alınarak genellenebilirlik kuramının, verilerin güvenirliğinde kullanılabileceği gibi puanlayıcılar arası güvenirliğin belirlenmesinde de tercih edilebileceği belirtilmiştir.

Lee ve Frisbie (1999), test setlerinden (testlets) oluşan testler üzerinde yapılan önceki çalışmalarda maddeye-dayalı güvenirlik tahminlerinin abartılı sonuçlara ulaşmış olabileceğini vurgulayarak, bu tür bir testin güvenirlik tahmininde genellenebilirlik teorisi yaklaşımının kullanılmasının gerekliliklerini ve uygunluğunu ortaya koymaya çalışmışlardır. Bu çalışmada, 1992 Bahar döneminde Iowa Temel Beceri Testleri (Iowa Tests of Basic Skills) ve Iowa Eğitimsel Gelişim Testleri (Iowa Tests of Educational Development)'nin 4., 8. ve 11. sınıflara uygulanmasından elde edilen puanlar kullanılmıştır. Testler, 3. sınıftan 12. sınıfa kadar tüm düzeyleri temsil edecek düzeyde olduğu için bu üç sınıf düzeyinde uygulama yapılması yeterli görülmüştür. Uygulamanın örneklem büyüklükleri, her test ve sınıf düzeyi için 3000 civarında öğrenciden oluşmuştur. Madde puanlarına dayalı Cronabach alfa kullanılarak elde edilen güvenirlik tahminlerinin genellenebilirlik teorisi yaklaşımı kullanılarak elde edilene göre 0.04 kadar daha fazla olduğunu bulmuşlardır. Değişen pasaj sayılarından ve sabit toplam madde sayısına dayalı elde edilen genellenebilirlik katsayılarının, sabit pasaj sayılarından ve sabit toplam madde sayısına dayalı elde edilen genellenebilirlik katsayılarından daha değişken olduğu sonucuna varmışlardır. Böylece, etkili ölçme çalışmalarının gerçekleşmesi için testlerdeki pasaj sayısını düzenlenmenin, her bir pasajdaki madde sayısını düzenlemekten daha verimli bir yol olacağını açıklamışlar, genellenebilirlik kuramının bu tür çalışmalarda sağladığı açıklayıcı bilgilerin önemine vurguda bulunmuşlardır.

McBee ve Barnes (1998), yaptıkları çalışmada, 8. Sınıf matematik performansının ölçülmesinin zamanlık sabitliği (temporal stability) ile görev-içi (intertask) tutarlılığını ve ölçme sonuçlarının genellenebilme kapasitesinin görevlerin benzerliklerinden nasıl etkilediğini araştırmışlardır. Karmaşık bir problem çözme alanını ölçmek üzere geliştirilen ikisi hayli benzer olan dört performans görevi analiz edilmiştir. Puanlayıcılar arası (interrater) güvenirlik katsayısı 0.80'lerde bulunurken, test tekrar-test güvenirlik

katsayısı 0.28 ile 0.54 arasında değişen hayli düşük değerler arasında kalmıştır. Genellenebilirlik çalışması sonuçları, görev ana etkisinin ve diğer değişkenlik kaynaklarıyla etkileşiminin oldukça büyük bir hata kaynağı oluşturduğunu ancak bu hata bileşenlerinin sadece benzer iki görev için analiz edildiğinde oldukça önemli bir düşüş meydana getirdiğini göstermiştir. Genellenebilirlik çalışması aynı zamanda istenilen düzeyde genellenebilirliğe ulaşmak için oldukça benzer görevlerde bile görev sayılarının oldukça fazla sayıda arttırılması gerektiğini göstermiştir. Araştırmanın sonuçları önceki çalışmalarda elde edilen bulgular ile uyumlu olarak, performansın ölçülmesindeki görevlerin kullanımında dikkatli olunması gerektiği konusunda vurguda bulunmuştur.

Kieffer (1998), genellenebilirlik kuramının niçin gerekli olduğunu, klasik test kuramının genelde yetersiz kaldığını; klasik test kuramına karşı genellenebilirlik kuramını kullanmanın yararlarını tartışmıştır. Çalışmada, genellenebilirlik kuramının klasik test kuramının bakış açısından yola çıkarak ve daha genişletilmiş bir hali olarak ortaya konulan bir kuram olduğu açıklanmıştır. Klasik test kuramında bir seferde sadece bir değişkenlik kaynağını (test tekrar-test, parelel formlar, iç tutarlılık vb...) araştırmak mümkün olurken, aynı anda birden fazla değişkenlik kaynağı ve bunlar arasındaki etkileşimin hesaplanması mümkün olmazken, genellenebilirlik kuramının çoklu ölçme hatası kaynaklarının ve bunlar arası etkileşimin büyüklüklerini aynı anda tek bir analizle hesaplayabilmeye imkanı verdiğini açıklamıştır. Çalışmada klasik test kuramının sınırlılıkları ve genellenebilirlik kuramını kullanmanın güçlü yanları açıklanmış ve bu açıklamalar verilen hipotetik bir örnekte elde edilen sonuçlarla desteklenmiştir.

Sub, Valiga ve Goa (1997), Akademik Danışmanlık Anketi (Academic Advising Survey, AAS)ne öğrencilerin verdiği puanların güvenirliğini genellenebilirlik kuramına dayanarak araştırmışlardır. Bu çalışmada, lise sonrası eğitim veren 15 farklı enstitüde görev yapan 150 danışmanın her birinin 10 öğrenciye akademik danışmanlıkta bulunarak, danışmaya katılan 1500 öğrencinin 18 maddelik AAS'ne verdikleri puanlara (birey:akademi:danışman)xmadde, rasgele etki (random effect) modeline dayalı G ve D çalışması uygulanmıştır. G çalışması sonucunda en büyük değişkenlik kaynağının hata (residual) varyansı olduğu görülmüş ve bunu takip eden diğer en büyük iki değişkenlik kaynağı sırasıyla öğrenciler (birey) ve (akademi:danışman) değişkenlik kaynağı olarak

bulunmuştur. D çalışması sonucunda, her bir danışmana düşen öğrenci sayısı 10'dan düşük olduğunda AAS'nin güvenirliği düşük (.63-.75), 10 ile 20 arasında olduğunda güvenirliğin orta düzeyde (.76-.87) ve öğrenci sayısı 20'nin üzerinde çıktığında ise güvenirliğin oldukça yüksek olduğu sonucuna varılmıştır. AAS'den elde edilen verilerin güvenirliği ayrıca klasik test kuramına dayalı olarak alfa katsayısıyla hesaplanarak 0.93 bulunmuştur. Ölçme sonuçları güvenirliğinin araştırılmasında genellenebilirlik kuramının kullanılmasının avantaj ve dezavantajları alfa katsayısı çalışması ile karşılaştırılarak tartışılmıştır. Çalışmada yer alan anketin ve bu tür anket çalışmalarının güvenirliğinin araştırılmasında, ölçme objesinin öğrenciler değil danışmanlar olduğu ve bu durumda danışmanların gerçek puan varyansının göz önüne alınması gerektiği vurgulanarak, alfa katsayısı ile elde edilen güvenirliğin yanlış yorumlanabileceği, bu durumda genellenebilirlik kuramına dayalı güvenirlik çalışması yapılmasının daha uygun ve doğru olacağı yorumu yapılmıştır.

Teresa (1997), araştırmasını 40'dan fazla ülkenin katıldığı 8. ve 9. sınıf öğrencilerine uygulanan Üçüncü Uluslararası Matematik ve Fen Çalışması (TIMMS) üzerinde yapmıştır. TIMSS'de yer alan 300 maddenin yaklaşık dörtte birini oluşturan maddeler, öğrencilerin kendi cevaplarını üretmelerini gerektiren açık-uçlu sorulardan oluşmaktadır. Araştırmasını bu açık uçlu sorulardan, 23'ü kısa cevaplı ve 8'i de açık uçlu olmak üzere toplam 31'i üzerinde yapmıştır. Bu soruların puanlanması 2-rakamlı tanımlayıcı koddan oluşan bir rubrikle yapılmıştır. Bu kodlamada yer alan ilk rakam, cevabın doğruluğunu belirlerken; ikinci rakam, cevabın genel yaklaşımını ya da yanlış yorumunu gösteren belirli bir tipini tanımlamıştır. Çalışmada, madde tipinin bir fonksiyonu olan puanlayıcı etkisine bağlı, bu açık-uçlu sorulardan elde edilen ülke düzeyindeki puanların ve öğrenci pualarının hata varyansı kaynaklarının göreceli katkısı tartışılmıştır. İngilizce konuşan yedi ülkenin her birine 50 öğrenciye uygulanmak üzere soru kitapçıkları verilerek, 350 öğrenciden elde edilen veriler analiz edilmiştir. Genellenebilirlik çalışmaları, puanlayıcılara bağlı etkiler ile ilişkili olarak puanların değişkenliğini belirlemiştir. Genellenebilirlik katsayıları, ülke-çaprazlama kodlama çalışmasından elde edilen verilere dayalı TIMMS'in açık uçlu sorularının puanlanmasında, ülkelerin ortalama puanlarına göre sıralamanın güvenirliğinin yüksek olduğunu göstermiştir. Belirli bir maddeden bireysel olarak her bir öğrencinin aldığı puan üzerinde yapılan genellenebilirlik bazı maddeler için daha az istikrarlı

bulunmuştur. Ancak TIMSS sonuçları ülke düzeyindeki ortalamalara bağlı olarak rapor edildiği için bu durumun endişe verici bir durum olmadığı vurgulanmıştır.

Moon (1995)'un yaptığı çalışmanın amacı, bir performansın ölçülmesinden elde edilen sınırlı sayıda gözlemin, gözlenemeyen olası durumlara uygun şekilde genellenip genellenemeyeceğini araştırmaktır. Öğrencilerin performanslarına ilişkin nasıl güvenilir ölçmeler yapılabileceği ve kağıt-kalem testlerine karşı performansın ölçülmesinden nasıl farklı bilgiler elde edilebileceği açıklanmıştır. Yapılan bu çalışmada, üniversite öğrencilerine istatistik dersine ilişkin hem performansın ölçülmesi ve hem de kağıt kalem testi birlikte uygulanmıştır. Performansın ölçülmesinin güvenirliğinin kestirimi ve desenin geliştirilmesi için genellenebilirlik çalışması yapılmıştır. Puanların güvenirliği klasik test kuramı ve genellenebilirlik kuramı ile analiz edilmiştir. Her iki ölçme aracına ilişkin göreceli bilgiler elde etmek için korelasyonel ve açıklayıcı (exploratory) faktör analizi yapılmıştır. Çalışmanın sonucunda, performansları puanlayan iki puanlayıcının önemli bir hata kaynağı olmadığı bulunmuştur. Ayrıca, öğrencilerin performanslarındaki tutarsızlığın başlıca kaynağının görevler olduğu gözlenmiş ve görev sayısının arttırılarak puanların güvenirliğinin arttırılabileceği sonucuna ulaşılmıştır. İki puanlayıcı tarafından elde edilen tüm puanlar arasındaki korelasyon ve G çalışmasının sonuçları, puanlayıcıların tutarlı bir şekilde öğrenci performanslarını değerlendirebildiklerini göstermiştir. Araştırmanın sonucunda puanlayıcı sayısının bir azaltılabileceği ve böylece genellenebilirlik çalışmasındaki desende yer alan puanlayıcı değişkenlik kaynağının çıkarılması önerilmektedir. İki ölçüm arasında oldukça yüksek korelasyon bulunmuştur. Faktör analizi sonuçları faktör yapıları ile madde ayırıcılıkları arasında bir ilişki olduğunu göstermiştir.

Shavelson, Webb ve Rowley (1989), ‘‘Genellenebilirlik Kuramı" başlığı altında yaptıkları çalışmada, genellenebilirlik kuramının, davranışların ölçülmesindeki güvenirlik çalışmaları için pratik ve esnek bir yaklaşım sağladığını göstermeye çalışmışlardır. Genellenebilirlik kuramının, klasik test kuramının daha genişletilmiş bir hali olduğunu göstermeye çalışmışlardır. Bu düşüncelerini desteklemek üzere genellenebilirlik kuramının klasik test kuramından farklı olarak (a) pek çok hata kaynağının büyüklüklerini tahmin edebildiği, (b) hem mutlak hem de göreceli kararlara varabilmek için yapılan ölçmelerde kullanılabilen bir model olduğu, (c) ölçme

amaçlarına göre farklı güvenirlik (genellenebilirlik) katsayıları sağladığı, (d) maliyet-etkili ölçmeler yapabilmek için başlıca hata kaynaklarını tespit edebildiği vurgulanmıştır. Bununla birlikte, genellenebilirlik kuramına ilişkin formüllerin ve kuramın geliştirilmesinin oldukça teknik olması dolayısıyla psikolojik araştırmalarda kolaylıkla kullanılamadığını vurgulamışlardır. Yapmış oldukları çalışmayla, genellenebilirlik kuramının uygulama alanının ne kadar geniş olduğunu ve geniş kitlelerce uygulanabileceğini göstermeye çalışmışlardır.

Yurtdışında genellenebilirlik kuramına dayalı yapılan çalışmalar, değişkenlik kaynaklarının doğru ve açık bir şekilde tanımlandığı durumlarda genellenebilirlik kuramının açıklayıcı sonuçlar verdiğini göstermektedir. Performansın ölçülmesine ilişkin çalışmalarda, genellenebilirlik kuramının klasik test kuramına göre daha kapsamlı ve aydınlatıcı bilgiler sunduğu, tercih edilebilir bir yöntem olduğu vurgulanmaktadır. Performansın ölçülmesinde genellenebilirlik kuramınına dayalı güvenirliğin araştırılmasıyla, her bir değişkenlik kaynağına ve bunlar arasındaki etkileşime ilişkin değişkenliği hesaplayabilmenin, hem göreceli hem de mutlak güvenirliği kestirebilmenin, farklı karar çalışmaları ile en uygun sayıda puanlayıcı, madde, ölçme durumu (occasion)na karar vermenin mümkün olduğu farklı araştırmalarda elde edilen ortak sonuçlar arasında yer almaktadır.

1.18.2.2. Çok Değişkenlik Kaynaklı Rasch Ölçme Modeline Dayalı Araştırmalar

Hetherman (2004), New York şehri eğitim birimi İngilizce kompozisyon yazma testinin etkinliğini çok değişkenlik kaynaklı Rasch ölçme modelini kullanarak araştırmıştır. Bu çalışmada, yazılı yoklamalarda (essay) çok boyutluluk fonksiyonu ve analitik puanlama yöntemi kullanılmış; özellikle farklı değişkenlik boyutlarına ilişkin değişimleri incelenmiştir. Bu amaç doğrultusunda çok değişkenlik kaynaklı Rasch ölçme modeli kullanılmıştır. Değerlendirmede yer alan tüm değişkenlik boyutları ve bunlar arasındaki etkileşimler incelenmiştir. Örneklemin 230 olduğu bu araştırmada, uygun olan hem 0-1 puanlama (dichotomous) hem çoklu puanlama (polytomous) Rasch modeli kullanılmıştır. Çalışma sonucunda: (a) Puanlayıcıların yazılı yoklamalarda kaydettikleri hatalar göz önüne alınmaksızın, uygulamanın yapıldığı bireylerin dil yeteneklerinin, (0, 1) yetenek sürekliliğinde iki farklı düzeyde oldukları anlaşılmıştır.

(b) Dil yeteneği bileşeninin güçlük düzeyinin -1.68 ile 1.33 arasında manidar bir değişim gösterdiği sonucuna varılmıştır. (c) Tümüyle puanlayıcılar arası güvenirliğin yüksek, standartlaştırılmış z puanlarının -1.48 ile 3.5 arasında olduğu gözlenmiştir. (d) Yazılı yoklamadaki farklı başlıkların (topics) güçlük düzeylerinin farklı olduğu sonucuna varılmıştır. (e) Yazılı yoklama sisteminin altında yatan varsayımların tutarsızlıkları ve çalışma bulgularının teorideki ve pratikteki yorumları üzerinde açıklamalarda bulunulmuştur. Ayrıca gelecek çalışmalar için önerilere yer verilmiştir.

Turner (2003), öğrencilerin resim yeteneklerini ölçmek için hazırlanan resim portfolyo değerlendirmesinin güvenirliğini çok değişkenlik kaynaklı Rasch ölçme modelinin kullanılarak incelemiştir. Bu çalışmada, fakültenin belirlediği, eğitim programının amaçlarıyla örtüşen 12 portfolyo projesinin 15 öğrenci tarafından tamamlanmasıyla elde edilen veriler kulllanılmıştır. Öğrencilerin tamamlamış olduğu bu projelerin tamamı 10 profesyonel resim tasarımcısı tarafından puanlanmış, elde edilen puanların güvenirliği çok değişkenlik kaynaklı Rasch ölçme modeli ile araştırılmıştır. Puanlama, projelerin farklı açılardan bir bütünlük oluşturduğu dört boyutun beş kategorili bir rubrikle değerlendirilmesiyle yapılmıştır. Her bir öğrencinin projesi eş olasılıkla seçilen üç puanlayıcı tarafından puanlanmıştır. Öğrencilerin yeteneği, projelerin ve puanlama boyutlarının güçlük düzeyleri ve puanlayıcı katılık/cömertlik seviyelerinin analizinde çok değişkenlik kaynaklı Rasch ölçme modeli kullanılmıştır. Çalışma sonucunda, ham puan ortalamaları esas alındığında, üç öğrencinin resim yeteneği kestirimlerin üstünde (over estimated) ve üç öğrencinin yeteneği de kestirimlerin altında (under estimated) bulunmuştur. Çok değişkenlik kaynaklı Rasch ölçme modeli, öğrenci puanlarını puanlayıcı katılık/cömertliklerine göre düzenlemiş ve böylece altı öğrencinin sıralaması değişmiştir. Puanlayıcı katılık/cömertlik boyutu için ayırma indeksi güvenirliği 0.93 ve puanlayıclar arası tutarlılık %43.3 olarak bulunmuştur. 0.73 ayırma indeksi güvenirliği ile puanlayıcıların ölçme boyutlarının farklılıklarını ayırmada yetersiz oldukları sonucuna varılmıştır. Ayrıca, projelerin ve puanlama boyutlarının güçlük düzeylerinin öğrenci yeteneklerinin belirlenmesine etkide bulunmadığı elde edilen veriler arasında yer almıştır.

Nakamura (2002), Japon sınıflarında öğretmen ve akranların İngilizce dilinde sözel sunum becerilerinin ölçülmesi sonuçlarının güvenirliğini çok değişkenlik kaynaklı

Rasch ölçme modeliyle incelemiştir. Çalışmada yer alan ölçme durumu, 12 yüksek lisans öğrencisinin (ölçme objesi) sınıftaki sözel sunumlarının (görev değişkenlik kaynağı), bir öğretmen ve dört akranları (puanlayıcı değişkenlik kaynağı) tarafından puanlanmasını kapsamaktadır. Çok değişkenlik kaynaklı Rasch ölçme modelinin, her bir değişkenlik kaynağı için elde edilen uyum istatistikleriyle uyumsuzluk (misfit) gösteren öğrenci, puanlayıcı ya da maddeleri belirlediği açıklanmıştır. Bu sonuçlara göre her bir değişkenlik kaynağında uyumsuzluk gösteren elemanların puanlamadan çıkarılarak tekrar yapılacak puanlamalarda ölçmenin daha güvenilir sonuçlar verecek şekilde nasıl geliştirilebileceğine ilişkin açıklamalara yer verilmiştir.

Nakamura (2000), iletişimsel dil testi sonuçlarına çok değişkenlik kaynaklı Rasch ölçme modelini uyguladığı tekniksel çalışmasında, iletişim dili testi verilerinin analizine çok değişkenlik kaynaklı Rasch ölçme modelinin nasıl uygulanabileceği ve bu modelin dil eğitimcileri açısından nasıl yararlı olabileceğini araştırmıştır. Çalışmasında, 20 maddelik konuşma (sohbet tarzı) testi ve 15 maddelik İngilizce sözel iletişim yeteneğini ölçen sosyo-linguistik testinin 30 üniversite öğrencisine uygulanarak altı İngilizce öğretmeninin puanlamasından elde edilen verileri kullanmıştır. Rasch (1960, 1980) tarafından 0-1 puanlama yapılan (dichotomous) test maddelerine uygulanan ve daha sonra da Andrich (1978) ve Masters (1982) tarafından tutum (attitude) anketlerinin, kısmi puanlama yapılan test maddeleri gibi sıralanmış kategorideki cevaplalarına genişletilen Rasch ölçme modelini Nakamura, üçüncü bir değişkenlik kaynağı olan puanlayıcıları da dahil ederek kullanmıştır. Böylece, iletişimsel dil testi verilerinin çok değişkenlik kaynaklı Rasch ölçme modeli analiz sonuçlarına dayalı olarak öğrenci, madde, puanlayıcı ve puan kategorisi değişkenlik kaynaklarına ilişkin ölçme raporlarını ve uyum istatistiklerini sunmuştur. Bunlara dayalı olarak, çok değişkenlik kaynaklı Rasch ölçme modeli analizinin (1) sadece genel (global) olarak değil, her bir değişkenlik boyutu (locally) olarak da yararlı bilgiler sunduğunu, (2) aynı ölçekte birbirleriyle göreceli olarak tüm değişkenlik kaynaklarının durumlarını sergilediğini, (3) puanlayıcılar için ayrı bir ölçme raporu vererek, araştırmacılara çok daha detaylı bilgi verebildiği yorumunda bulunmuş ve veriler üzerinde bu yorumlarını örneklendirerek açıklamıştır.

Engelhard (1996), performansın ölçülmesinde puanlayıcıların doğruluğunu araştırmış ve puanlayıcıların doğruluğunu değerlendirmek için yeni bir yöntem açıklamıştır. Uygulayıcı, puanlayıcılardan elde edilen puanlar ile bir grup benchmark, örnek ya da asıl dayanak noktası olan performanslara ilişkin bir uzman panelinden elde edilen puanlar arasındaki uyumu "doğruluk" olarak tanımlamıştır. Çalışmasında, Rasch Modeli'nin bir uzantısı olan FACETS modelini sunmuştur. FACETS yöntemini, 20 uygulayıcı puanlayıcı ve bir uzman panelin 373 benchmark kağıdını puanlamasıyla elde edilen puanlarla örneklendirmiştir. Çalışmada kullanılan veriler, Georgia'daki "High School Graduation Writing Test"inden elde edilmiştir. Veriler, puanlayıcı doğruluğunda istatistiksel olarak manidar bir farklılığın olduğunu göstermiştir. Veriler aynı zamanda, benchmark kağıtlarının diğerlerinden çok daha kolaylıkla doğru olabileceğini göstermektedir. Engelhard, ayrıca doğruluğun değerlendirilmesinde kullanılan benchmarkların farklı alt testlerinde puanlayıcıların doğruluklarının nasıl değişmez olabileceğine dair küçük bir örnek vermiştir.

Brown, W. L., O'Gorman, K. ve Du, Y. (1995), matematik performansının ölçülmesinin geçerlik ve güvenirliğini araştırdıkları çalışmalarında; Minneapolis devlet okullarında uygulanan matematikte problem çözme performansının ölçülmesinin psikometrik özelliklerini incelemişler, ölçmenin geçerliği ve puanlayıcılar arası güvenirliği ile puanlama güvenirliğine odaklanmışlardır. Çalışmada 5. sınıftaki 1300 öğrenci tarafından yanıtlanan 2600 kağıt FACETS modeli kullanılarak analiz edilmiştir. FACETS genel bir ölçekte performansın ölçülmesinin her bir değişkenlik kaynağını ölçeklendiren ve ölçek üzerinde her bir değişkenlik kaynağındaki elemanları diğer değişkenlik kaynağındaki elemanlarla karşılaştırmaya imkan veren, bir parametreli madde tepki kuramının ya da Rasch Modelinin genişletilmiş bir halidir. Klasik test kuramı ile FACETS kullanımı değerlendirmeciye puanlayıcılar arası güvenirlik, puanlayıcılar arası uyum ve geçerliliği test etme olanağı sağlar. Bu çalışmanın sonuçları matematikte problem çözme performansının ölçülmesinin geçerlik ve güvenirliğini desteklemiştir. Çalışma ayrıca, güvenilir ve geçerli bir performans ölçme oluşturmada FACETS ve klasik test kuramının birlikte kullanılmasının avantajlarını da ortaya koymuştur. Klasik test kuramı ile birlikte FACETS kullanarak, matematikte problem çözme performansının ölçülmesinde puanlayıcılar arası uyum, puanlayıcılar arası güvenirlik ve geçerlik test edilmiştir.

Yapılan araştırmalar sonucunda, performansın ölçülmesine ilişkin çalışmalarda, genellenebilirlik kuramıyla birden fazla değişkenlik kaynağı birlikte göz önünde bulundurularak tek bir güvenirlik kestiriminde bulunulurken; çok değişkenlik kaynaklı Rasch ölçme modeline dayalı çalışmalar, her bir değişkenlik kaynağına ilişkin farklı farklı güvenirlik kestirimlerinde bulunmanın mümkün olduğunu göstermektedir. Yurt dışında yapılan çalışmalardan elde edilen bir diğer sonuç; çok değişkenlik kaynaklı Rasch ölçme modeliyle her bir değişkenlik kaynağındaki elemanlara ilişkin uyumsuzlukların tespit edilmesiyle, bu elemanların çalışmadan çıkarılarak daha güvenilir sonuçlara ulaşılabileceğidir.

1.18.2.3. Genellenebilirlik Kuramı ve Çok Değişkenlik Kaynaklı Rasch Ölçme Modeline Dayalı Araştırmalar

MacMillian (2000), klasik test kuramı, genellenebilirlik kuramı ve çok değişkenlik kaynaklı Rasch ölçme modelini, geniş bir veri setinde puanlayıcılar arası değişkenliğin belirlenmesi ve bu değişkeliğin düzeltilmesi için karşılaştırmıştır. Bu çalışmada, 4930 öğrencinin katıldığı İngilizce sınavının, 70 puanlayıcıdan oluşan havuzdan çekilen 3 puanlayıcının puanlama sonuçlarının değişkenliğini araştırılmıştır. Puanlayıcılar arası değişkenlik klasik test kuramı, genllenebilirlik kuramı ve çok değişkenlik kaynaklı Rasch ölçme modeli yaklaşımları kullanarak analiz edilmiştir. Klasik test kuramı ve çok değişkenlik kaynaklı Rasch ölçme modeli yaklaşımlarına göre puanlayıcılar arası önemli farklılıklar olduğu belirtilmiştir. Çok değişkenlik kaynaklı Rasch ölçme modeli, klasik test kuramı yaklaşımıyla yapılan analizin ortaya çıkardığı puanlayıcılar arası farklılıktan daha büyük farklılık göstermiştir. Aksine, genellenebilirlik kuramından elde edilen puanlayıcı varyans bileşenleri ve Rasch histogramları puanlayıcılar arası değişimin az olduğu izlenimini vermiştir. Klasik test kuramı ve çok değişkenlik kaynaklı Rasch ölçme modeli düzeltme yöntemlerinin her ikisi de öğrencilerin %50'den daha fazlası için farklı puanlar üretmiş, fakat öğrencilerin %75'i, her bir düzeltmeden sonra aynı sonuçları vermiştir.

Tüm kuramlara dayalı güvenirliğin kestirimine ilişkin yapılan bu araştırmada her bir kuramın avantaj ve dezavantajları açıklanmaya çalışılmıştır. Klasik test kuramının, performans değerlendirme çalışmalarında güvenirliğin kestirimindeki sınırlılıkları yanı

sıra bu kurama dayalı yöntemleri kullanmanın kolaylığı ve aşinalığının bir avantaj olduğu açıklanmıştır. Genellenebilirlik kuramı performansın ölçülmesine ilişkin çalışmalarda tercih edilebilir bir yöntem olarak sunulmuş, kullanılmasının ve yorumlanmasının karmaşıklığı kuramın dezavantajı olarak verilmiştir. Çok değişkenlik Rasch ölçme modeli, tüm değişkenlik kaynaklarının tek bir ölçekte yer alarak değerlendirilebildiği, her bir değişkenlik kaynağına ilişkin farklı farklı güvenirliğin hesaplanmasının mümkün olduğu ancak az sayıda verinin bulunduğu durumlarda kullanılmasının mümkün olmadığı, kullanılmasının ve yorumlanmasının karmaşık olduğu vurgulanmıştır.

BÖLÜM II

YÖNTEM

Bu bölümde, araştırmanın türü, verilerin toplandığı grup, ölçme araçları ve verilerin analizine verilmektedir.

2.1. ARAŞTIRMANIN TÜRÜ

Araştırma, klasik test kuramı (KTK), genellenebilirlik (G) kuramı ve çok değişkenlik kaynaklı Rasch ölçme modelinin (ÇDKRÖM) uygulanmasına, birbirlerine olan üstünlük ve sınırlılıklarının ortaya konulmasına, yaklaşımlardan hangisinin daha çok bilgi sağladığının belirlenmesine dayanmaktadır. 1999 yılında yapılan TIMSS standart testinde yer alan performansın ölçülmesine uygun olan maddeler, 2007 yılında 8. ve 9. sınıf öğrencilerine uygulanarak; klasik test kuramı, genellenebilirlik kuramı ve çok değişkenlik kaynaklı Rasch ölçme modeli ile özelliklerinin belirlenmesine, durum saptamaya yönelik olduğundan bu yönüyle çalışma betimsel bir araştırma niteliği taşımaktadır. Öte yandan, klasik test kuramı, genellenebilirlik kuramı ve çok değişkenlik kaynaklı Rasch ölçme modelinin kıyaslanmış olması bakımından temel bir araştırmadır.

2.2. ÇALIŞMA GRUBU

Araştırmanın çalışma grubu, 2007 yılı bahar döneminde öğrenim gören, aşağıdaki Tablo 4'te verilen toplam 203 öğrenciden oluşmaktadır.

Tablo 4. Çalışma Grubunda Yer Alan Okullar ve Öğrenci Sayıları

OKULLAR	Öğrenci Sayıları
Elmadağ Lisesi 9. sınıf	35 öğrenci
Fethiye K. M. Anadolu Lisesi 9. sınıf	46 öğrenci
Samanyolu Koleji 8. sınıf	35 öğrenci
Uluğbey Anadolu Lisesi 9. sınıf	30 öğrenci
Ulus End. Meslek Lisesi 9. sınıf	28 öğrenci
Yavuz S. S. Anadolu Lisesi 9. sınıf	29 öğrenci
TOPLAM	203 öğrenci

Araştırmanın çalışma grubunu, Ankara ili merkezi, ilçe merkezi, genel lise, meslek lisesi, anadolu lisesi ve özel lise gibi farklı yerleşim yerlerinde ve okul türlerinde öğrenim gören 8. ve 9. sınıf öğrencileri oluşturmaktadır. Matematik performansının ölçülmesine ilişkin uygulama ders saati içinde yapıldığından, öğrencilerin seçimini belirleyen asıl unsur dersin öğretmeninin ve öğrencilerin gönüllülüğü olmuştur.

2.3. VERİ TOPLAMA ARACI

2007 yılı bahar döneminde yukardaki tabloda verilen okullarda öğrenim gören 8. ve 9. sınıf öğrencilerine, TIMSS'de yer alan 24 performans ölçme maddesi uygulanmıştır. Uygulanan maddeler Ek 1'de verilmiştir. Performansın ölçülmesinde yer alan 24 maddenin 6'sının, yapılan açıklayıcı ve doğrulayıcı faktör analizi sonucunda, iki farklı faktörde de göreceli olarak yüksek yük değerine sahip oldukları görüldüğünden çıkarılmasının uygun olduğuna karar verilmiştir (Büyüköztürk, 2006). Çalışmada yer alan tüm analizler kalan 18 madde üzerinden gerçekleştirilmiştir.

Matematik Performans Rubriği: Araştırmanın verileri Ek 2'de verilen bütünsel dereceli puanlama anahtarı kullanılarak elde edilmiştir. Her bir matematik performans maddesi 0-5 puan üzerinden ölçülmüştür. Bu puanlamanın özeti aşağıda verildiği şekildedir:

0 puan: Boş
1 puan: Sadece doğru cevap
2 puan: Uygun olmayan cevap
3 puan: Başlar ama problemi sonlandıramaz
4 puan: Tatmin edici cevap
5 puan: Örnek cevap

Matematik performansının ölçülmesinde yer alan 24 madde için hesaplanan Cronbach alfa iç tutarlılık katsayısı 0.92; faktör analizi sonucu, analizlerde kullanılmak üzere seçilen 18 madde üzerinden hesaplanan Cronbach alfa iç tutarlılık katsayısı ise 0.91 olarak bulunmuştur.

Verilerin elde edilmesine ilişkin süreç aşağıdaki gibi gerçekleşmiştir:

1) Öncelikle 1999 yılında uygulanmış TIMSS'de yer alan açık- uçlu 24 matematik sorusu seçilmiş ve yeni bir soru kitapçığı oluşturulmuştur. Oluşturulan soru kitapçığı, anlam ve dil bilgisi açısından iki matematik uzmanı tarafından tekrar gözden geçirilmiştir. Her bir madde tüm öğrencilerin rahatlıkla okuyabilecekleri punto büyüklüğünde yazılmış, oluşturulan soru kitapçıkları öğrencilerin cevaplarını rahatlıkla yazabilecekleri şekilde düzenlenmiştir. Matematik performansının ölçülmesi için düzenlenmiş olan 24 maddeden oluşan soru kitapçığı Ek 1'de verilmiştir.
2) Soru kitapçıklarının hazırlığı ve çoğaltma işlemi bittikten sonra, matematik performansının ölçülmesine ilişkin soru kitapçıkları çalışma grubundaki okullarda uygulanmıştır. Uygulama öncesi öğrencilere gerekli açıklamalar araştırmacı tarafından yapılmış, öğrencilerin varsa soruları araştırmacı tarafından cevaplandırılmıştır.
3) Çalışma grubundaki okullarda yapılan uygulamalar tamamlandıktan sonra matematik uzmanı olan ve gönüllü oldukları için seçilen 4 puanlayıcıya, araştırmacı tarafından hazırlanan dereceli puanlama anahtarı (Wiggins, 1998) ve puanların yazılacağı "puanlama tabloları" verilmiştir. Hazırlanan dereceli puanlama anahtarı ve puanlama tabloları için ayrıca uzman görüşü alınmıştır. Puanlama tablolarında, her bir öğrencinin cevap kağıdına verilen, okulunu simgeleyen bir harf ve her bir öğrenci için ayrı ayrı verilmiş bir numaradan oluşan kodlar bulunmaktadır. Diğer bir deyişle, her bir öğrencinin, isminin yer almadığı, bir harf ve bir sayıdan oluşan kod numarasının karşısına her bir madde için verdiği cevaba karşılık gelen puanların yazılacağı tablolar oluşturulmuştur. Bu puanlama tablosuna ilişkin bir örnek Ek 3'te verilmiştir.
4) Puanlayıcılara, dereceli puanlama anahtarının ve puanlama tablolarının nasıl kullanılacağı araştırmacı tarafından açıklanmış, puanlayıcılara puanlama için gerekli ve yeterli süre verilmiştir.
5) Sürecin sonunda, tüm puanlayıcılardan elde edilen puanlama tablolarındaki veriler araştırmacı tarafından bilgisayar ortamında Excel tablolarına aktarılarak, veriler araştırma için gerekli analizlerin yapılmasına hazır hale getirilmiştir.

2.4. VERİLERİN ANALİZİ

Bu çalışmada, matematik performansını ölçebilecek dört matematik uzmanı görev almıştır. Bu puanlayıcılar tamamen birbirlerinden bağımsız puanlama yapmışlardır. Dört puanlayıcının 24 maddeye verdikleri puanlar arasındaki korelasyon değerleri aşağıdaki Tablo 5'te verilmiştir.

Tablo 5. Dört Puanlayıcının 24 Maddeye Verdikleri Puanlar Arasındaki Korelasyon Katsayıları

	1.Puanlayıcı	2.Puanlayıcı	3.Puanlayıcı	4.Puanlayıcı
1.Puanlayıcı	-	0.946	0.903	0.958
2.Puanlayıcı		-	0.956	0.972
3.Puanlayıcı			-	0.935

Verilerin analizi üç aşamada gerçekleştirilmiştir. İlk aşamada, dört puanlayıcıya göre ayrı ayrı, 203 öğrencinin 24 maddeye verdikleri cevaplardan elde edilen puanların betimsel istatistikleri hesaplanmıştır.

Genellenebilirlik kuramı, klasik test kuramındaki güvenirlik ile geçerlik arasındaki farkın, yapılacak uygun gözlemler ile ortadan kaldırılabileceğini savunmaktadır. Klasik test kuramındaki güvenirlik ve geçerlik yerine G (genellenebilirlik) ve Phi (güvenirlik) katsayılarının kullanımı yeterli olmaktadır. Bu sebeple, genellenebilirlik kuramındaki güvenirlik ile geçerliğin birlikte kullanımının daha iyi anlaşılması açısından ikinci aşamada matematik performansınının ölçülmesi güvenirliğinden önce yapı geçerliği incelenmiştir. Bu amaçla, yine dört farklı puanlayıcı için faktör analizi çalışması yapılmıştır. Açklayıcı faktör analizi çalışmasıyla belirlenen faktör yapıları doğrulayıcı faktör analiziyle sınanmıştır. Açıklayıcı faktör analizi sonuçları dört puanlayıcıdan elde edilen puanlar için yapılmış benzer sonuçlar ürettiği gözlenmiştir. Ancak çalışmadaki tüm puanlayıcıların betimsel istatistikleri farklılık gösterdiğinden (bknz. Tablo 6) çalışmada, puanlayıcıların her bir maddeye verdikleri puanların ortalamaları üzerinden yapılan açıklayıcı ve doğrulayıcı faktör analizi sonuçlarına yer verilmiştir.

Son aşamada ise veriler araştırmanın alt problemlerine göre sırasıyla klasik test kuramı, genellenebilirlik kuramı ve çok değişkenlik kaynaklı Rasch ölçme modeli için kullanılan desenlere uygun olarak analiz edilmiştir.

Cronbach α, açıklayıcı faktör analizi ve Kendall'ın uyum katsayısı için SPSS 14 programı kullanılmıştır. Doğrulayıcı farktör analizi için Lisrel 8.7 programından yararlanılmıştır. Genellenebilirlik kuramına göre ana ve ortak etkiler için varyans değerlerinin kestirilmesi, puanların güvenirliği için G ve Φ(Phi) katsayılarının hesaplanmasında SPSS (14.0) (Musquash ve O'Connor, 2006) ve ÇKDRÖM analizleri için FACETS (Linacre 2007) programları kullanılmıştır.

BÖLÜM III

BULGULAR VE YORUM

Bu bölümde, araştırmanın problem cümlesi ve alt problemlerin sırasına göre elde edilen bulgular ve bulgulara ilişkin yapılan yorumlar yer almıştır.

Bulguların verilmesinden önce, dört puanlayıcıya göre ayrı ayrı, 203 öğrencinin 24 maddeye verdikleri cevaplardan elde edilen puanların betimsel istatistikleri Tablo 6'da verilmiştir.

Tablo 6. Öğrenci Puanlarının Dört Puanlayacıya İlişkin Betimsel İstatistikleri (N=203)

İstatistikler	Puanlayıcılar			
	1	2	3	4
Ortalama	57	67	68	67
Ortanca	59	73	76	74
Tepe Değer	81	88	83	90
Std. Sapma	20.4	20.4	20.5	21.7
Varyans	417.17	417.05	420.89	471.84
Çarpıklık	-0.53	-1.08	-1.28	-0.96
Basıklık	-0.55	0.59	0.96	0.14
α güvenirliği	0.91	0.91	0.92	0.92

Tablo 6'da görüldüğü gibi, 203 öğrencinin 24 maddeden 0 - 5 puan ölçeği üzerinden aldıkları puanlara ilişkin en yüksek ortalama 68 ile 3. puanlayıcıya ve en düşük ortalama 57 ile 1. puanlayıcıya aittir. Diğer iki puanlayıcıya ait ortalama değerler birbirine eşit olup 67'dir. Her bir puanlayıcıya ilişkin ortanca aritmetik ortalamadan büyüktür ve puanlayıcılara ait puanlar sola çarpık dağılım göstermektedir. Bu durum, tüm puanlayıcılar için çarpıklık katsayısının negatif çıkmasıyla da görülebilmektedir. Basıklık katsayılarına bakıldığında, 1. puanlayıcının verdiği puanların basıklık katsayısı 0'dan küçük bir değer olup puanlar normalden daha basık, diğer puanlayıcıların verdiği puanların basıklık katsayısı 0'dan büyük bir değer olup puanlar normalden daha sivri dağılım göstermektedirler. Tüm puanlayıcıların verdikleri puanların Cronbach alfa (α) güvenirlik katsayıları birbirine eşit denecek kadar yakın ve oldukça yüksek değerlerdir.

3.1. ALT PROBLEM I: KLASİK TEST KURAMINA GÖRE MATEMATİK PERFORMANSININ ÖLÇÜLMESİNİN İNCELENMESİ

3.1.1. Matematik Performansının Ölçülmesinde Yer Alan Maddelerin Yapı Geçerliği Çalışması

Matematik performansının ölçülmesinin geçerliğinin test edilmesi için yapı geçerliğine bakılmıştır. Bu amaçla, dört puanlayıcıya ilişkin elde edilen betimsel istatistikler farklı sonuçlar gösterdiğinden (Tablo 6), faktör analizi çalışması tüm puanlayıcıların verdikleri puanların ortalamaları üzerinden yapılmıştır. Açıklayıcı faktör analizi çalışmasıyla belirlenen faktör yapıları ayrıca doğrulayıcı faktör analiziyle sınanmıştır (Şekil 2).

Faktör analizi çalışmasında tüm maddelerin tek bir faktörde toplandığı gözlenmiş ve bu faktör "matematik başarısı" olarak tanımlanmıştır. Matematik performansının ölçülmesinde yer alan 24 maddeden altısı (1., 2., 5., 10., 16. ve 22. maddeler) iki farktörde de yüksek yük değerleri verdiğinden bu maddeler çıkarılarak, bundan sonraki tüm analizler kalan 18 madde üzerinden yapılmıştır. Ek 4'te 24 maddenin tümüne ilişkin faktör analizi yük değerleri verilmiştir. Matematik performansının ölçülmesinin dört puanlayıcıya ilişkin açıklayıcı faktör analizi sonuçları Tablo 7'de verilmiştir.

Tablo 7. Matematik Performansının Ölçülmesine İlişkin Dört Puanlayıcıdan Elde Edilen Puanların Açıklayıcı Faktör Analizi Sonuçları

İstatistikler	Puanlayıcılar			
	1	2	3	4
1'den büyük öz değerler	3	3	3	3
Üç faktörün varyans açık.	%52	%53	%56	%54
Bir faktörün varyans açık.	%39	%40	%42	%42

Faktör analizinin uygulanabilmesi için örneklem büyüklüğünün korelasyon güvenirliğini sağlayacak ölçüde olması gerekmektedir. Dolayısıyla örneklemden elde edilebilen verilerin, faktör analizi yapmak için uygun olup olmadığının sınanması için Kaiser-Meyer-Olkin (KMO) testi yapılmış ve KMO=0.92 bulunmuştur (Büyüköztürk, 2006).

Faktör analizinin uygulanıp uygulanamaycağına karar verebilmek için kullanılan bir diğer test de Barlett ölçüsüdür. Barlette ölçüsü, değişkenlere (maddeler) ilişkin korelasyon matrisinin birim matrise eşit olduğu kabulune dayalı yokluk hipotezini test eder. Faktör analizinin uygulanabilmesi için değişkenler arasında bir ilişki olması gerekir ve eğer korelasyon matrisi birim matrise eşit olursa tüm korelasyon katsayıları sıfır olacaktır. Barlett testi ile bu durum sınanmaktadır. Barlett testi sonucunda p değeri α=0.05 değerinin altında çıkarsa, yokluk hipotezi reddedilerek, korelasyon matrisinin birim matrisinden farklı olduğu yorumu yapılır. Böylece faktör analizi için istenilen, değişkenler arasında ilişki olduğu sonucuna varılabilir (Field, 2005). Bu çalışmada, matematik performansının ölçülmesiyle elde edilen puanlara Barlett testi uygulanmış ve 1397.637 olarak 0.01 düzeyinde anlamlı bulunmuştur. Elde edilen bu sonuç, verilere faktör analizinin uygulanabileceğini göstermektedir.

Açıklayıcı faktör analizi sonucunda, faktör yük değerleri 0.40'ın üzerindekiler alınmış ve bu ölçüte göre 18 maddenin tamamının birinci faktörde bulunduğu gözlenmiştir (Tablo 7). Birinci faktördeki yük değerleri 0.44 ile 0.78 arasında değişmektedir. Temel bileşenler analizinde birinci faktörün açıkladığı toplam varyans %39.26'dır. Bu durum matematik performansının ölçülmesinin genel bir faktöre sahip olduğunu göstermektedir. Matematik performansının ölçülmesine ilişkin temel bileşenler analizi sonuçları Tablo 8'de verilmiştir.

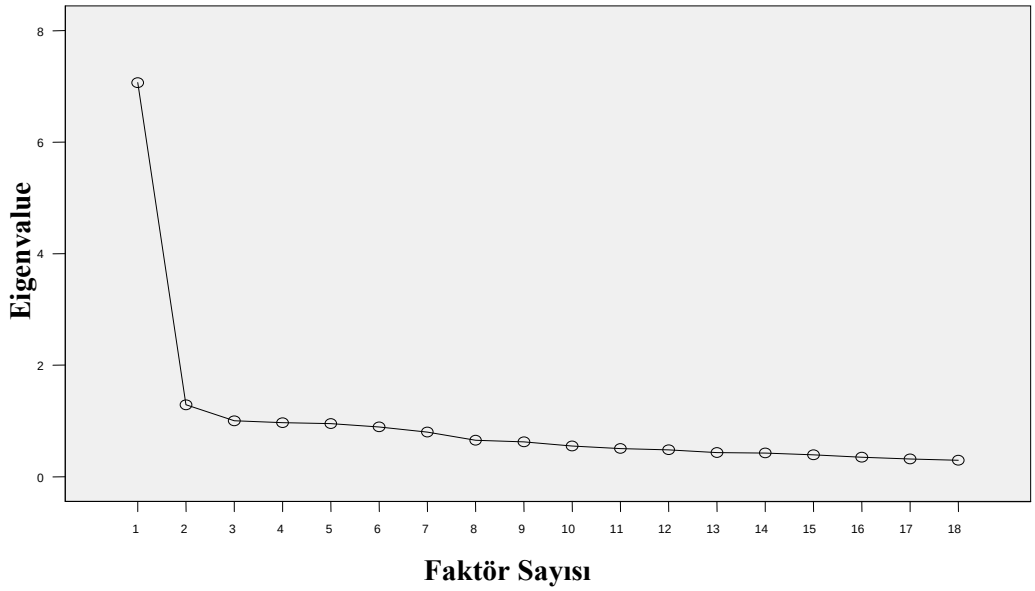
Tablo 8. Matematik Performansının Ölçülmesine İlişkin Dört Puanlayıcının Puanları Ortalamalarına Ait Özdeğerler ve Açıklanan Varyans Yüzdeleri

Faktörler	Özdeğerler	Açıklanan Varyans Yüzdeleri	Açıklanan Yığılmalı Varyans Yüzdeleri
1	7.07	39.26	39.26
2	1.29	7.16	46.42
3	1.00	5.57	52.00

Tablo 8'e göre, matematik performansının ölçülmesine ilişkin temel bileşenler analizinde, 18 maddenin öz değerleri 1.00'den büyük olan üç faktör bulunmaktadır. Dolayısıyla matematik performansının ölçülmesi en fazla üç faktörlü kabul edilebilir.

Üç faktörün birlikte açıkladıkları toplam varyans %52'dir. Temel bileşenler analizine göre, birinci faktörün öz değeri 7.07, açıkladığı varyans %39.26, ikinci faktörün öz değeri 1.29 açıkladığı varyans %7.16, üçüncü faktörün öz değeri 1.00 açıkladığı varyans %5.57'dir.

Öz değerlere göre çizilen grafikten de (Şekil 2) anlaşılacağı üzere önemli faktör sayısı 1 olarak görülmektedir. Grafikte, ikinci faktöre kadar oldukça hızlı (sert), ikinci ve sonraki faktörlerde grafiğin genel gidişi yatay olup, hızlı (sert) bir düşüş eğilimi gözlenmemektedir. Kısacası, ikinci ve sonraki faktörlerin varyansa katkıları birbirine oldukça yakın ve düşüktür. Buna göre tüm maddeleri tek bir faktör altında toplamanın uygun olacağı sonucuna varılmıştır.



Şekil 2. Matematik Performansınının Ölçülmesiyle Elde Edilen Verilerin Açıklayıcı Faktör Analizinde Öz Değerler Grafiği

Matematik performansının ölçülmesinde yer alan maddelerin temel bileşenler analizi sonuçları, tek faktördeki yükleri ve madde-test korelasyon katsayıları Tablo 9'da verilmiştir. Maddeler üç faktördeki yüklerine göre en yüksek faktör yükünden başlayarak verilmiştir.

Matematik performansının ölçülmesinde yer alan 18 maddeden aynı yapıyı ölçen maddelerin belirlenmesinde, maddelerin bulundukları faktördeki yüklerinin yüksek

olması ve maddelerin tek bir faktörde yüksek, diğer faktörlerde ise düşük yük değerlerine sahip olması kriteri göz önünde bulundurulmuştur. Maddelerin faktör yüklerinin en az 0.30 olması ve tek faktörde yer alması (maddenin iki ayrı faktörde yüksek faktör yüküne sahip olması durumunda farkın en az 0.10 olması) ölçütlerine uyulmuştur (Büyüköztürk, 2006).

Yukarıda belirtilenler ışığında, matematik performansının ölçülmesinde yer alan 18 maddenin yükleri 0.54 ile 0.78 arasında değişmektedir. Tüm maddeler birinci faktörde toplanmıştır. Böylece matematik performansının ölçülmesinin tek boyutlu olduğu ve bu boyutun da "matematik başarısı" olarak isimlendirilebileceği söylenebilir.

Tablo 9'da "Matematik başarısı" boyutu altında yer alan 18 maddenin madde-test korelasyonlarının 0.49 ile 0.76 arasında olduğu görülmektedir.

Tablo 9. Matematik Performansının Ölçülmesinde Yer alan 18 Maddenin Temel Bileşenler Analizi Sonuçları, Birinci Faktördeki Yük Değerleri ve Madde-Test Korelasyon Sonuçları

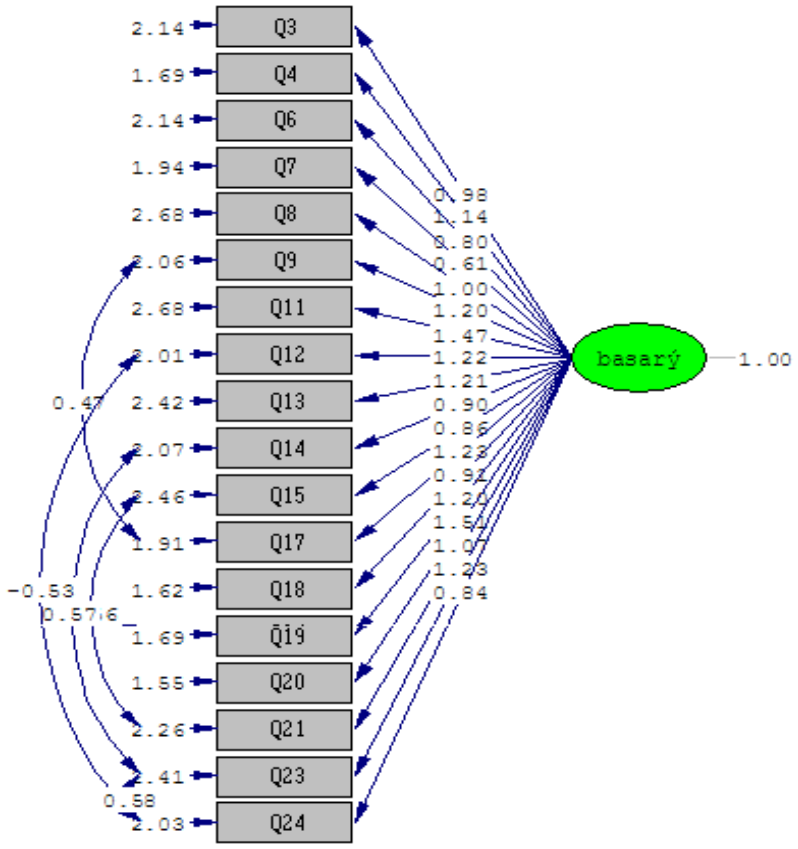
Madde	Faktör Ortak Varyansı	1. Faktör	Madde-Test Korelasyonu
3	0.55	0.59	0.54
4	0.55	0.67	0.61
6	0.37	0.52	**0.49**
7	0.36	0.54	0.50
8	0.49	0.56	0.53
9	0.48	0.67	0.67
11	0.51	0.68	0.62
12	0.60	0.67	0.62
13	0.59	0.63	0.64
14	0.34	0.58	0.59
15	0.61	0.54	0.57
17	0.55	0.70	0.72
18	0.39	0.61	0.57
19	0.51	0.70	0.67
20	0.65	0.78	**0.76**
21	0.51	0.63	0.60
23	0.63	0.68	0.70
24	0.69	0.55	0.58
18 madde için Cronbach α katsayısı: 0.92			

Matematik performansının ölçülmesine ilişkin doğrulayıcı faktör analizinin teorik modeli şekil 3'te görüldüğü gibi elde edilmiştir. Modelin bir bütün olarak değerlendirilmesi için üretilen uyum iyiliği değerleri incelendiğinde, modelin uyum iyiliği ölçütlerine göre iyi değerler verdiği ve bu şekliyle kolaylıkla kabul edilebileceği görülmektedir.

Uygulanan analiz sonucunda uyum indeksleri χ^2=230.60 (sd=130, p<0.001), (χ^2/sd)=1.77, RMSEA=0.062, RMR=0.17, standardize edilmiş RMR=0.053, GFI=0.99, AGFI=0.98, CFI=1.00, NFI=1.00 ve NNFI=1.04 olarak elde edilmiştir. Ki-kare değerinin (230.60) serbestlik derecesine (130) oranı 2'nin altındadır ve bu iyi uyum olduğunun bir göstergesidir.

Benzer şekilde, RMSEA ve CFI değerleri de iyi bir uyum olduğuna işaret etmektedir. Bununla birlikte, modele dair AIC (312.60) ve CAIC (489.45) değerlerinin, bağımsız modeli (sırasıyla,4109.41 ve 4187.05) ve doymuş model (sırasıyla 342.00 ve 1079.56) değerlerinden daha düşük oldukları görülmektedir. Bu değerler, uyum iyiliği için gerekli değerlere ulaşıldığını göstermektedir.

Sonuç olarak, açıklayıcı ve doğrulayıcı faktör analizi çalışmalarına göre matematik performansının ölçülmesinde yer alan maddelerin "matematik başarısı" olarak adlandırılan tek bir boyut altında toplandıkları yorumuna ulaşılabilir.



Şekil 3. Doğrulayıcı Faktör Analizi Sonucunda Elde Edilen Matematik Performansının Ölçülmesinin Teorik Modeli

3.1.2. Klasik Test Kuramına Göre Matematik Performansının Ölçülmesinin İç Tutarlılığı

Öğrencilerin matematik performansınının ölçülmesinden aldıkları puanların dört puanlayıcıya göre güvenirlik analizleri yapılmıştır. Performans ölçmelerinin yapısına ilişkin temel varsayımlardan biri de ölçmede yer alan her bir maddenin gözlenmek istenen performansla ilişkili olmasıdır. Bu her bir maddenin istenilen performansı ölçebildiğini ifade eder. Bu sebeple, performansın ölçülmesinde yer alan maddelerin iç tutarlılıklarının belirlenmesi gerekmektedir. Bu amaca yönelik olarak da her bir puanlayıcının verdiği puanlar için Cronbach α güvenirlik katsayısı hesaplanmıştır. Birbiriyle yüksek ilişki gösteren maddelerin α güvenirlik katsayısının da yüksek olması beklenmiştir. Matematik performansının ölçülmesine ilişkin Cronbach α güvenirlik katsayıları birinci ve ikinci puanlayıcılar için 0.91 ve üçüncü ve dördüncü puanlayıcılar

için 0.92 olarak bulunmuştur. Elde edilen bu katsayılara göre, matematik performansının ölçülmesinde yer alan maddelerin matematik başarısını birbirleriyle tutarlı bir şekilde ölçtükleri yorumu yapılabilir.

3.1.3. Klasik Test Kuramına Göre Matematik Performansının Ölçülmesinde Dört Farklı Puanlayıcının Puanları Arasındaki Tutarlılık Düzeyi

Matematik performansının ölçülmesinde, dört farklı puanlayıcının aynı koşullar altında her bir öğrenciyi puanlamasına ilişkin elde edilen puanlar arasındaki tutarlılık derecesi, parametrik olmayan istatistiksel bir teknik olan Kendall'ın uyum katsayısı (Kendall's Concordance Coefficient) ile analiz edilmiş ve sonuç olarak 18 madde için 0.52 olarak bulunmuştur (X^2=315.16, sd=3, p=0.00<0.05). Elde edilen bu sonuca göre puanlayıcılar arası uyum olduğu yorumu yapılabilir (Howell, 2002). Ayrıca, her bir puanlayıcının verdiği puanlar ile diğer bir puanlayıcının verdiği puanlar arasındaki korelasyon katsayıları hesaplanmıştır. Puanlayıcıların 18 madde üzerinden verdikleri puanlar arasındaki korelasyon değerleri Tablo 10'da verilmiştir.

Tablo 10. Dört Puanlayıcının 18 Maddeye Verdikleri Puanlar Arasındaki Korelasyon Katsayıları

	1.Puanlayıcı	2.Puanlayıcı	3.Puanlayıcı	4.Puanlayıcı
1.Puanlayıcı	-	0.943	0.902	0.955
2.Puanlayıcı		-	0.956	0.974
3.Puanlayıcı			-	0.940

Tablo 10'da görüldüğü gibi dört puanlayıcının 18 maddeye verdiği puanlar arasındaki korelasyon katsayıları 0.90 ile 0.97 arasında değişen oldukça yüksek değerlere sahiptir. Elde edilen bu değerler puanlayıcılar arası uyumun olduğu yorumunu destekler niteliktedir.

Ancak puanlayıcılar arası korelasyon katsayıları ortalamalardan bağımsız olarak hesaplandığı için, puanlayıcıların verdikleri puanların ortalamaları arasındaki farklılıkları ortaya koyamaz. Bu nedenle puanlayıcıların puanları ortalamaları arasındaki farklılığın test edilmesi önerilmektedir (Goodwin, 2001). Bu çalışmada yer alan dört puanlayıcının 203 öğrenciye verdiği puanların ortalamaları arasındaki farklılık, ilişkili örneklem üzerinde tek değişkenli varyans analizi ile test edilmiş ve istatistiksel olarak anlamlı bir fark olduğu sonucuna varılmıştır (F=13.801, p=0.00< α=0.05). Bu sonuç

üzerine yapılan post-hoc çalışmasıyla puanlayıcıların puanları ortalamalarının ikili karşılaştırılması sonucunda birinci puanlayıcı hariç tüm puanlayıcıların verdikleri puanların ortalamaları arasında manidar bir farklılık görülmemiştir. Sadece birinci puanlayıcının puanları ortalaması diğer tüm puanlayıcıların puanları ortalamasından istatistiksel olarak anlamlı bir farklılık göstermiştir.

3.2. ALT PROBLEM II.: GENELLENEBİLİRLİK KURAMINA GÖRE MATEMATİK PERFORMANSININ DEĞERLENDİRİLMESİNİN İNCELENMESİ

3.2.1. Matematik Performansının Değerlendirilmesinin Kestirilen Genellenebilirlik Kuramı Parametreleri

"Matematik başarısı" olarak adlandırılan tek boyutta toplanan matematik performansının ölçülmesinin G çalışması ile elde edilen varyanslarını ve varyans yüzdelerini hesaplamak amacıyla, 2007 yılında yapılan TIMSS'den alınan 18 açık-uçlu maddeye tümüyle çaprazlanmış b x g x p modeli uygulanmıştır. Ölçmenin uygulandığı 203 öğrenci, 4 puanlayıcı ve tek boyutlu ölçmeden (görev) oluşan orjinal verilerle tek değişkenli modelle yapılan G çalışması için; kestirilen varyans bileşenleri ve toplam varyansı açıklama yüzdeleri b, g ve p ana etkileri ile bg, bp, gp ve bgp ortak etkileri aşağıdaki Tablo 11'de verilmiştir. Tablo 11'de verilen tek değişkenli G çalışması sonucunda matematik performansınının ölçülmesinin tek boyutu için kestirilen varyans ve toplam varyansı açıklama oranları incelendiğinde, birey (b) ana etkisi için kestirilen varyans bileşeninin (1.160) toplam varyansın % 32.6'sını açıkladığı görülmektedir. Tek değişkenli modelle bireyler için kestirilen varyans bileşenini, toplam varyans içinde en yüksek ikinci sırada paya sahip varyans bileşenidir. Genellenebilirlik çalışmalarında, birey ana etkisi evren puanı varyansı olarak değerlendirilir ve ölçülen özellik açısından bireyler arası farklılaşmayı ifade eder (Shavelson ve Webb, 1991; Brennan 2001).

Bireyler için kestirilen varyansın toplam varyans içindeki oranının en büyük olması istenilen bir durumdur. Bu durum, ölçme ile elde edilen boyutta bireyler arası farklılıkların ortaya çıkarılabildiğinin bir göstergesidir.

Tablo 11. Matematik Performansının Ölçülmesinin Tek Değişkenli G Çalışması Sonucunda Ölçmenin Kestirilen Varyansları ve Toplam Varyansı Açıklama Oranları

Varyans Kaynağı	Sd	Toplam Kareler	Kareler Ortalaması	Varyans	%
b	202	18407.652	91.127	1.160	32.6
g	17	2397.441	141.026	0.159	4.5
p	3	838.823	297.608	0.075	2.1
bg	3434	23194.197	6.754	1.553	43.6
bp	606	837.150	1.381	0.047	1.3
gp	51	292.933	5.744	0.026	0.7
bgp	10.302	5604.094	0.544	0.544	15.3
Toplam					100

Görev (g) ana etkisi için tek değişkenli modelle yapılan G çalışmasından kestirilen varyans bileşeni (0.159) toplam varyansın % 4.5'unu açıklamaktadır. Görev ana etkisinin varyans bileşeni büyüklüğünün, diğer ana etkiler içinde ikinci sırada yer aldığı görülmektedir. Puanlayıcı ana etkisi için kestirilen varyans bileşeni (0.075) toplam varyansın % 2.1'ini açıklayarak, ana etkiler içinde en düşük varyans bileşeni olmaktadır. Puanlayıcı ana etkisinin G çalışması ile kestirilen varyans oranının düşük çıkması, puanlayıcıların tüm bireyler için yaptıkları puanlamalar arasında çokça bir farklılık bulunmadığının, puanlamalar arasında bir tutarlılığın söz konusu olduğunun göstergesi olabilir.

Birey x görev (bg) ortak etkisi (1.553) toplam varyansın %43.6'sını açıklamaktadır. Birey x görev ortak etkisi tek değişkenli analizle elde edilen en büyük varyanstır. Bu durum, tek değişkenli analizle bu ölçme için birey x görev ortak etkisinden kaynaklanan farklılığın büyük olduğunu, belli bireylerin bağıl durumlarının bir görevden diğerine çok farklılaştığını göstermektedir. Birey x puanlayıcı (bp) ortak etkisi (0.047) toplam varyansın % 1.3'ünü açıklamaktadır. Birey x puanlayıcı ortak etkisi tek değişkenli analizle elde edilen en küçük ikinci varyansa sahiptir. Buna göre, tek değişkenli analizle bu ölçme için birey x puanlayıcı ortak etkisinden kaynaklanan farklılığın küçük olduğu, belli bireylerin bağıl durumlarının bir puanlayıcıdan diğerine çok değişmediği yorumu yapılabilir. Görev x puanlayıcı (gp) ortak etkisi (0.026) toplam varyansın % 0.7'sini açıklamaktadır. Görev x puanlayıcı ortak etkisi tek değişkenli analizle elde edilen en küçük varyanstır. Bu sonuca göre, tek değişkenli analizle bu ölçme için görev x puanlayıcı ortak etkisinden kaynaklanan farklılığın neredeyse olmadığı, belli görevlere

verilen puanların puanlayıcıdan puanlayıcıya değişmediği yorumunda bulunulabilir. Birey x görev x puanlayıcı (artık) ortak etkisi varyans bileşeninin de (0.544) toplam varyansı açıklama oranı % 15.3 olarak görülmektedir. Bu oran tablodaki üçüncü en büyük değerdir. Birey x görev x puanlayıcı (artık) varyansın büyük olması; birey, puanlayıcı ve görev ortak etkisi ve /veya tesadüfi hataların büyük olabileceğinin bir göstergesi olabilir.

3.2.2. Orjinal Puanlayıcı ve Madde Sayıları ile Puanlayıcı Sayılarının Bir Arttırılıp Bir Azaltılması ve Madde Sayılarının Üçte Bir Arttırılıp Azaltılması Senaryoları Sonucunda K Çalışmasıyla Kestirilen, G ve Phi Katsayıları

2007 yılında uygulanan matematik performansının ölçülmesinin "matematik başarısı" tek boyutunda yer alan 18 madde sayısına ve bu sayının üçte biri kadar maddelerin azaltılıp arttırılmasına, 4 puanlayıcı sayısına ve bu sayının birer arttırılıp azaltılmasına göre düzenlenen senaryolar için yapılan K çalışması sonucunda kestirilen mutlak ve bağıl hata varyansları Tablo 12 ve Şekil 4'te verilerek açıklanmıştır.

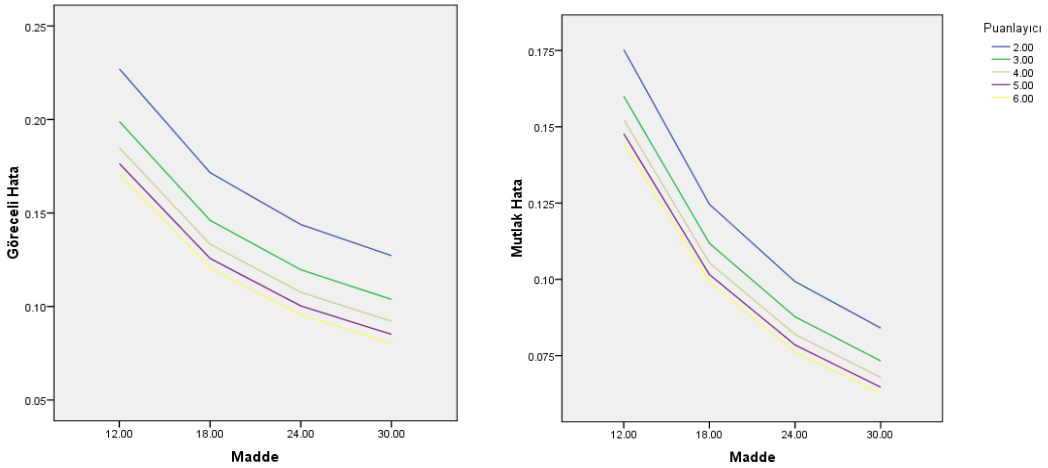
Tablo 12. Matematik Performansınının Ölçülmesine İlişkin K Çalışması ile Madde ve Puanlayıcı Sayıları Senaryolarına Göre Mutlak ve Bağıl Hata Varyansları

Madde Sayıları	Puanlayıcı Sayıları									
	2		3		4		5		6	
	Mutlak	*Bağıl*	Mutlak	*Bağıl*	Mutlak	*Bağıl*	Mutlak	*Bağıl*	Mutlak	*Bağıl*
12	0.227	*0.175*	0.199	*0.160*	0.185	*0.152*	0.171	*0.148*	0.171	*0.145*
18	0.172	*0.125*	0.146	*0.112*	**0.133**	***0.105***	0.126	*0.102*	0.121	*0.099*
24	0.144	*0.099*	0.120	*0.088*	0.108	*0.082*	0.100	*0.079*	0.095	*0.076*
30	0.127	*0.084*	0.104	*0.073*	0.092	*0.068*	0.085	*0.065*	0.080	*0.063*

Tablo 12'de görüldüğü gibi, matematik performansının ölçülmesinin tek boyutu altında yer alan 18 madde sayısına ve bu maddeleri puanlayayan 4 puanlayıcıya göre elde edilen mutlak hata varyansı 0.133 ve bağıl hata varyansı da 0.105 olarak kestirilmiştir. Buradan da görüleceği gibi, asıl uygulamadan (18 madde ve 4 puanlayıcı) elde edilen puanların mutlak hata varyansı bağıl hata varyansından daha yüksektir. Madde sayısı sabit tutularak (18) puanlayıcı sayısının bir azaltılması (3) durumunda kestirilen mutlak

hata varyansı 0.146 ve bağıl hata varyansı 0.112, puanlayıcı sayısının bir arttırılması (5) durumunda ise mutlak hata varyansı 0.126 ve bağıl hata varyansı 0.102 olarak kestirilmiştir. Puanlayıcı sayısı sabit tutularak (4) madde sayısının üçte bir azaltılması (12) durumunda kestirilen mutlak hata varyansı 0.185 ve bağıl hata varyansı 0.152, madde sayısının üçte bir arttırılması (24) durumunda ise mutlak hata varyansı 0.108 ve bağıl hata varyansı 0.082 olarak hesaplanmıştır. Buradan da görüleceği üzere, puanlayıcı sayısı sabit tutularak madde sayısının üçte bir artması durumunda mutlak hatadaki azalma 0.024 iken, madde sayısı sabit tutularak puanlayıcı sayısının bir arttırılması durumunda mutlak hatadaki azalma 0.007 olmaktadır. Benzer durum, bağıl hata için de incelenirse, puanlayıcı sayısının sabit tutularak madde sayısının üçte bir arttırılmasında azalma 0.023 iken, madde sayısı sabit tutularak puanlayıcı sayısının bir arttırılmasındaki azalma ise 0.003 olarak gözlenmektedir. Tablo 12 ve Şekil 4'te görüleceği üzere, puanlayıcı varyansının toplam varyansa oranının (% 0.075) daha önceki bölümde açıklandığı gibi sıfıra oldukça yakın olması sebebiyle, puanlayıcı sayısındaki artış ya da azalma mutlak ve bağıl hata varyansını fazlaca etkilememektedir. Benzer şekilde, puanlayıcı sayısı sabit tutularak madde sayısının arttırılması durumunda mutlak ve bağıl hata varyanslarının, uygulamadaki puanlayıcı ve madde sayısında kestirilen mutlak ve bağıl hata varyanslarından çok da farklı olmadığı gözlenmektedir. Örneğin; uygulamadaki madde (18) ve puanlayıcı (4) sayılarının ikisi birden arttırıldığında (madde sayısı 24 ve puanlayıcı sayısı 5 olduğunda) mutlak hata varyansındaki azalma 0.033 ve bağıl hata varyansındaki azalma ise 0.026 olarak görülmektedir. Puanlayıcı ve madde sayılarını arttırmanın, hata varyanslarında elde edilen bu kadar küçük azalmalara sebep olması dolayısıyla, emek ve zaman açısından fazlaca ekonomik olmadığı yorumu yapılabilir.

Tablo 12'de, uygulamadaki madde ve puanlayıcı sayılarının azalması durumunda mutlak ve bağıl hata varyanslarının giderek arttığı, arttırılması durumunda ise mutlak ve bağıl hata varyanslarının giderek azaldığı açıkça görülmektedir. Tüm madde ve puanlayıcı sayılarına göre, mutlak hata varyansı bağıl hata varyansından daha yüksek değerlere sahiptir. Tablo 12'deki değerlere göre, hem mutlak hem de bağıl hata varyansı için de madde sayılarını üçte bir arttırmanın, puanlayıcı sayısını bir arttırmaya göre daha büyük düşüşe sebep olduğu söylenebilir.



Şekil 4. Matematik Performansınının Ölçülmesine İlişkin K Çalışması ile Madde ve Puanlayıcı Sayıları Senaryolarına Göre Mutlak ve Bağıl Hata Varyansları

2007 yılında uygulanan matematik performansının ölçülmesinin "matematik başarısı" tek boyutunda yer alan 18 madde sayısına ve bu sayının üçte biri kadar maddelerin azaltılıp arttırılmasına, 4 puanlayıcı sayısına ve bu sayının birer arttırılıp azaltılmasına göre düzenlenen senaryolar için yapılan K çalışması sonucunda G (genellenebilirlik) ve Phi "*Φ*" (güvenirlik) katsayılarına ilişkin değerler Tablo 13 ve Şekil 5'te verilerek açıklanmıştır.

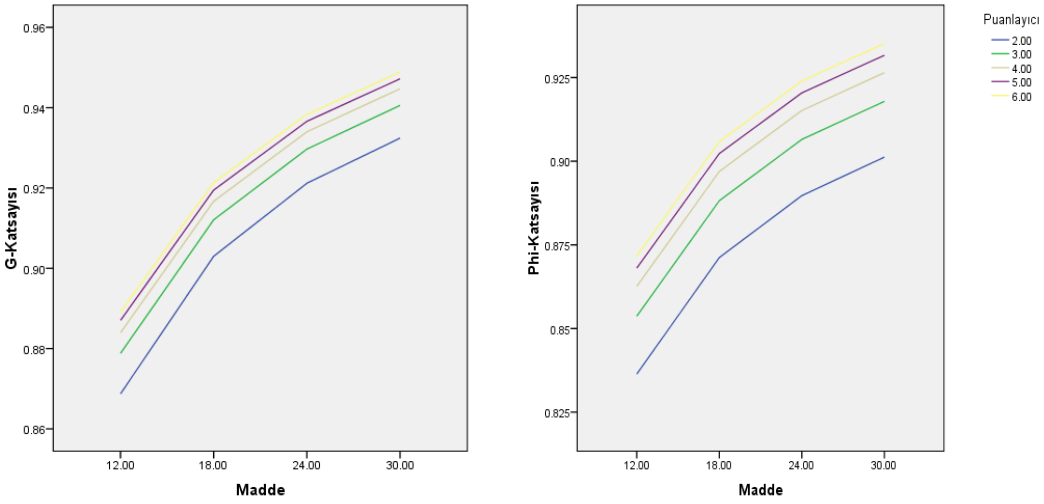
Tablo 13. Matematik Performansının Ölçülmesine İlişkin K Çalışması ile Madde ve Puanlayıcı Sayıları Senaryolarına Göre G ve Phi Katsayıları

Madde Sayıları	Puanlayıcı Sayıları									
	2		3		4		5		6	
	G	*Φ*	G	*Φ*	G	*Φ*	G	*Φ*	G	*Φ*
12	0.869	*0.836*	0.879	*0.854*	0.884	*0.863*	0.887	*0.868*	0.889	*0.872*
18	0.903	*0.871*	0.912	*0.888*	**0.917**	***0.897***	0.919	*0.902*	0.921	*0.906*
24	0.921	*0.890*	0.930	*0.907*	0.934	*0.915*	0.937	*0.920*	0.938	*0.924*
30	0.932	*0.901*	0.941	*0.918*	0.945	*0.926*	0.947	*0.932*	0.949	*0.935*

Tablo 13'te görüldüğü gibi, matematik performansının ölçülmesinin tek boyutu altında yer alan uygulamadaki 18 madde sayısına ve bu maddeleri puanlayayan 4 puanlayıcıya göre elde edilen G katsayısı 0.917 ve *Φ* katsayısı da 0.897 olarak kestirilmiştir. Buradan görüleceği gibi, uygulamadaki durumda değerlendirmeden elde edilen puanların G katsayısı *Φ* katsayısından daha yüksektir. Madde sayısı sabit tutularak (18) puanlayıcı

sayısının bir azaltılması (3) durumunda kestirilen G katsayısı 0.912 ve Φ katsayısı 0.888, puanlayıcı sayısının bir arttırılması (5) durumunda ise G katsayısı 0.919 ve Φ katsayısı 0.902 olarak kestirilmiştir. Puanlayıcı sayısı sabit tutularak (4) madde sayısının üçte bir azaltılması (12) durumunda kestirilen G katsayısı 0.884 ve Φ katsayısı 0.863, madde sayısının üçte bir arttırılması (24) durumunda ise G katsayısı 0.934 ve Φ katsayısı 0.915 olarak hesaplanmıştır. Buradan da görüleceği üzere, puanlayıcı sayısı sabit tutularak madde sayısının üçte bir artması durumunda G katsayısındaki artma 0.017 iken, madde sayısı sabit tutularak puanlayıcı sayısının bir arttırılması durumunda G katsayısındaki artma 0.002 olmaktadır. Benzer durum, Φ katsayısı için de incelenirse, puanlayıcı sayısının sabit tutularak madde sayısının üçte bir arttırılmasında meydana gelen artma 0.018 iken, madde sayısı sabit tutularak puanlayıcı sayısının bir arttırılmasındaki artma ise 0.005 olarak gözlenmektedir. Tablo 13 ve Şekil 5'te görüleceği üzere, puanlayıcı varyansının toplam varyansa oranının (% 0.075) daha önceki bölümde açıklandığı gibi sıfıra oldukça yakın olması sebebiyle, puanlayıcı sayısındaki artış ya da azalma G ve Φ katsayılarını fazlaca etkilememektedir. Benzer şekilde, puanlayıcı sayısı sabit tutularak madde sayısının arttırılması durumunda da G ve Φ katsayılarının, uygulamadaki puanlayıcı ve madde sayısında kestirilen G ve Φ katsayılarından çok da farklı olmadığı gözlenmektedir. Örneğin; uygulamadaki madde (18) ve puanlayıcı (4) sayılarının ikisi birden arttırıldığında (madde sayısı 24 ve puanlayıcı sayısı 5 olduğunda) G katsayısındaki artma 0.020 ve Φ katsayısındaki artma ise 0.023 olarak görülmektedir. Puanlayıcı ve madde sayılarını arttırmanın, G ve Φ katsayılarında elde edilen bu kadar küçük artmalara sebep olması dolayısıyla, emek ve zaman açısından fazlaca ekonomik olmadığı yorumu yapılabilir.

Tablo 13'te uygulamadaki madde ve puanlayıcı sayılarının azalması durumunda G ve Φ katsayılarının giderek azaldığı, arttırılması durumunda ise G ve Φ katsayılarının giderek arttığı açıkça görülmektedir. Tüm madde ve puanlayıcı sayılarına göre, G katsayısı Φ katsayısından daha yüksek değerlere sahiptir. Tablo 13'teki değerlere göre, hem G katsayısı hem de Φ katsayısı için de madde sayılarını üçte bir arttırmanın, puanlayıcı sayısını bir arttırmaya göre daha büyük artışa sebep olduğu söylenebilir.



Şekil 5. Matematik Performansınının Ölçülmesine İlişkin K Çalışması ile Madde ve Puanlayıcı Sayıları Senaryolarına Göre G ve *Φ* Katsayıları

3.3. ALT PROBLEM III.: ÇOK DEĞİŞKENLİK KAYNAKLI RASCH ÖLÇME MODELİNE GÖRE MATEMATİK PERFORMANSININ ÖLÇÜLMESİNİN İNCELENMESİ

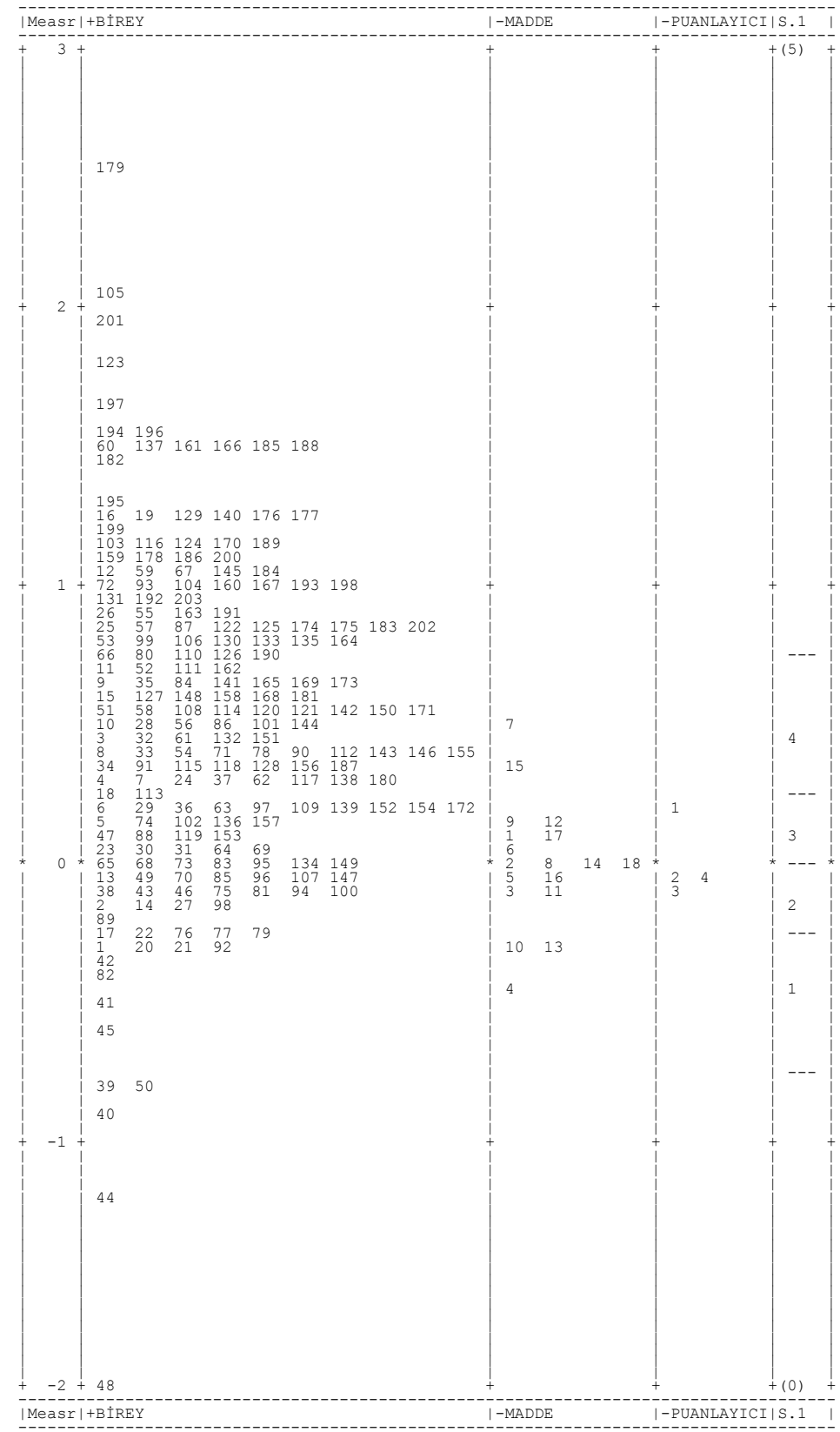
3.3.1. Çok Değişkenlik Kaynaklı Rasch Ölçme Modeli Analizleri

2007 yılında uygulanan matematik performansının ölçülmesine ilişkin 18 maddenin çok değişkenlik kaynaklı Rasch ölçme modeli analizleri için FACET bilgisayar programında, öğrenciler, maddeler ve puanlayıcılardan oluşan üç değişken kaynaklı desen kullanılmıştır. Çalışmanın bu aşamasında, analiz sonucunda elde edilen model veri uygunluğu, öğrencilerin yetenekleri ve uygunluk istatistikleri, puanlayıcıların katılık/cömertlik istatistikleri, madde güçlükleri ve uygunluk istatistikleri sırasıyla verilmiştir. Analiz sonuçları öğrenci örneklemi üzerinden genellenmektedir.

Çok değişkenlik kaynaklı Rasch ölçme modeli, öğrencilere ilişkin ölçümlerin (puanların) kestiriminin, puanlayıcıların katılık/cömertlik ve maddelerin güçlük düzeylerinin kestirimlerinden bağımsız olarak yapılmasına imkan verir. Bu durum da öğrencilerin puanlarının kestirimlerinin testten bağımsız olmasını sağlar. Rasch ölçme modeli aynı zamanda "bölgesel bağımsızlık" (=local independent) özelliğine de sahiptir. Bölgesel bağımsızlık; ölçmelerin istatistiksel kestirimlerinin, hangi puanlaycının hangi öğrenciye hangi maddede puanlama yaptığından bağımsız gerçekleşmesidir (Linacre, 1994). Rasch ölçme modelinin (madde tepki kuramını temel aldığından) kullanmanın

bir güçlü yanı da, öğrencilerin yetenek kestirimlerinin belirli bir testin maddelerinden bağımsız ve test maddelerinin parametre kestirimlerinin de bu maddelerin uygulandığı belirli bir grup öğrenciden bağımsız olarak yapılabilmesidir (Crocker ve Algina, 1986).

Çok değişkenlik kaynaklı Rasch ölçme modeli için FACET analizi sonuçlarında, her bir değişkenlik kaynağına ilişkin ölçümler "logit cetveli (haritası)" adı verilen tek bir doğrusal ölçekte gösterilebilir. Her bir değişkenlik kaynağı, öğrencilerin performanslarının gözlenen değerlerinden ve "0" orjin merkezinden itibaren ölçeklendirilir. Ancak, sadece öğrenci değişkenlik kaynağı serbestçe ölçeklendirilmektedir (Turner, 2003). Matematik performansının ölçülmesinden elde edilen verilere ilişkin üç değişkenlik kaynağı için elde edilen logit cetveli (haritası) Şekil 6'da gösterilmiştir. Değişkenlik kaynaklarına ilişkin ölçmeler bu haritada logit olarak ifade edilir. "logit" gözlenen puanların logoritmik dönüşümlerine dayalı eşit aralıklı bir metriği ifade etmektedir (Wright ve Stone, 1979). Bu metrikte, bağımsız değişkenler öğrenciler, puanlayıcılar ve maddeler olmak üzere değişkenlik kaynakları iken bağımlı değişken başarma olasılığı oranının logoritmik dönüşümü olmaktadır. (Hetherman, 2004). Şekil 6'da görülebileceği gibi değişkenlik kaynaklarının tümü aynı ölçek üzerinde ölçeklendirilebilmektedir. Şekil 6'da verilen cetvel beş sütundan oluşmaktadır. İlk sütun, modelde yer alan her bir değişkenlik kaynağının değerlendirildiği logit ölçeğini göstermektedir. İkinci sütunda her bir öğrenci numarasıyla gösterilerek, yetenek (θ) puanlarına göre sıralanmış olarak yer almaktadır. Bu sütunda, daha yüksek logit değere sahip öğrenciler daha yüksek yetenek düzeyinde bulunacak şekilde ölçeklenmiştir. Buna göre en yüksek yetenek düzeyindeki öğrenci 179 numaralı (2.50 logit birimde), en düşük yetenek düzeyindeki öğrenci de 48 numaralı öğrenci (-4.50 logit birimde) olarak görülmektedir. Bundan sonraki sütunda maddelerin güçlük derecelerine göre logit ölçeğindeki sıralanışı yer almaktadır. Ölçeğin en üstünde en zor soru, en altında da en kolay soru olacak şekilde maddeler güçlük derecelerine göre sıralanmışlardır. Buna göre, en zor soru 7 numaralı, en kolay soru da 4 numaralı soru olarak görülmektedir. Bir sonraki sütun ise, puanlayıcıların ölçeklendirilmesini göstermektedir.



Şekil 6. Matematik Performansınının Ölçülmesinin Çok Değişkenlik Kaynaklı Rasch Ölçme Modeline Göre Logit Haritası

Bu sütundaki değerler, logit puanı yüksek puanlayıcıların puanlamada daha katı, düşük puanlayıcıların ise puanlamada daha cömert olduğu şeklinde yorumlanmaktadır. Buna göre, 1 numaralı puanlayıcı en katı, 3 numaralı puanlayıcı ise en cömert puanlayıcı olarak görülmektedir. En son sütun ise puanlama kategorilerinin logit ölçekteki yerlerini göstermektedir. Bu sütun incelendiğinde puanlama kategorisinde 0,1 ve 2 puanları logit ölçekte 0 biriminin altında 3, 4 ve 5 puanları ise logit ölçekte 0 birimin üstünde yer almaktadır. Buna göre öğrencilerin büyük bir kısmının puanlama ölçeğindeki 3 puanının, logit birimde 0 birimin üstünde yer aldıkları gözlenmektedir. Öğrenci, puanlayıcı ve madde değişkenlik kaynaklarına ilişkin daha ayrıntılı bilgiler sırasıyla alt başlıklar olarak aşağıda verilmiştir.

3.3.2. Çok Değişkenlik Kaynaklı Rasch Ölçme Modelinde Model Veri Uygunluğu

Verilerin modelle olan uyumu için standartlaştırılmış artık değerlerin (residuals) yaklaşık en fazla %5'i ±2'nin dışında ve yaklaşık en fazla %1'inin de ±3'ün dışında olması gerekmektedir (Linacre, 2007). Matematik performansının ölçülmesinden elde edilen toplam 14.616 verinin standart artık değerlerinin 244'ü (%1.6) ±2'nin dışında ve 88'i (%0.6) de ±3'ün dışında yer almıştır. Buna göre, 2007 yılında uygulanana matematik performansının ölçülmesinin FACET analizi için model veri uyumunu sağladığı yorumu yapılabilir. Verilere ilişkin, tüm değişkenlik boyutlarının birlikte yer aldığı logit haritası Şekil 6'da verilmiştir.

3.3.3. Öğrencilerin Yetenekleri ve Uygunluk İstatistikleri

Matematik performansının ölçüldüğü toplam 203 öğrencinin logit değerleri -4.50 ile 2.50 arasında değişmektedir. Bu logit değerlerin ortalaması 0.51 ve standart sapması 0.11'dir. Ayırma indeksi (G) 4.39 ve güvenirliği 0.95 olarak yeterince yüksek bulunmuştur. Ayırma indeksi, değişkenlik kaynaklarına ilişkin elde edilen ölçmelerin sürekliliği boyunca ne ölçüde birbirlerinden ayrıldıklarını veren bir istatistiktir (Turner, 2003). Öğrenci değişkenlik boyutu için ayırma indeksi güvenirliği; KR-20, Cronbach alfa ve genellenebilirlik katsayısına benzer yorumlanabildiği için, testin iç tutarlılık anlamında güvenirliğinin yüksek olduğu sonucuna varılabilir (Nakamura, 2000).

Sabit etki (fixed effect) hipotezi X^2 testi ile reddedilmiştir (X^2=4388.9, serbestlik derecesi=201, p=0.00 < 0.05). Bu sonuca göre, öğrencilerin matematik performansının ölçülmesine dayalı olarak farklı düzeylere ayrılabildikleri yorumu yapılabilir. Öğrencilerin, istatistiksel olarak anlamlı kaç farklı yetenek düzeyine ayrılabildiği (4G+1)/3 formülü kullanılarak hesaplanabilir (Lee ve Kantor, 2003; Hetherman, 2004). Bu formüle dayalı olarak, 2007 yılında matematik performansının ölçülmesinin uygulandığı 203 öğrenci yaklaşık altı farklı yetenek düzeyine ayrılabilmektedir. Rasgele etki (random effect) hipotezi X^2 testi ile red edilmemiştir (X^2=187.1, serbestlik derecesi=200, p=0.73 > 0.05). Bu sonuca göre,verilerin normal dağılım gösterdiği ve verilerin modele uyumlu olduğu yourumu yapılabilir.

Hem iç uyum (infit) istatistikleri (ortalama=1, std.sapma=0.3) hem de dış uyum (outfit) istatistikleri (ortalama=1, std.sapma=0.4) verilerin modele çok iyi uyum gösterdiğine işaret etmektedir. 203 öğrencinin %96'sının (sadece 9 öğrenci hariç) uyum içi kareler ortalamaları ve %91'inin (sadece 18 öğrenci hariç) uyum dışı kareler ortalamalarının 0.5 ile 1.5 arasında olduğu gözlenmektedir. Bu değerler, verilerin ölçme için uygun (verimli) olduğuna işaret etmektedir (Linacre, 2007).

3.3.4. Puanlayıcıların Katılık/Cömertlik İstatistikleri

Matematik performansının ölçülmesinde yer alan dört puanlayıcının, çok değişkenlik kaynaklı Rasch ölçme modeli analizi ile elde edilen özellikleri ve istatistikleri Tablo 14'te verilmiştir.

Tablo 14'te dört puanlayıcının puanlamasındaki katılık/cömertliklerine ilişkin bulgular yer almaktadır. Tabloda puanlayıcıların logit değerlerinin -0.10 ile 0.22 arasında değiştiği görülmektedir. Bu logit değerlerin ortalaması 0.0 ve standart sapması 0.13'tür. Puanlayıcılar arasında en cömert puanlayıcı üçüncü puanlayıcı (-0.10) ve en katı puanlayıcı da birinci puanlayıcı (0.22) olarak gözlenmektedir. Ayırma indeksine bakıldığında 10.59 gibi oldukça yüksek bir değerle karşılaşılmaktadır. Bu değer puanlayıcılar arası farklılık olduğunun bir göstergesidir (Nakamura, 2002).

Tablo 14. Matematik Performansının Ölçülmesinde Yer Alan Puanlayıcıların Ölçme Raporu

			Puanlayıcı Katılığı		Uygunluk İçi		Uygunluk Dışı	
Puanlayıcı No	Puanlayıcı Ort.	Puanlayıcı Toplam r	Logit Ölçüsü	S.H.	Kareler Ort.	Z Std.	Kareler Ort.	Z Std.
1	3.2	3.60	0.22	0.01	0.9	-3	1.1	1
2	3.8	4.25	-0.06	0.01	1.0	0	1.0	0
3	3.8	4.32	-0.10	0.01	1.0	0	1.0	0
4	3.7	4.23	-0.05	0.01	1.1	3	0.9	-1
Ortalama	3.6	4.10	0.0	0.01	1.0	0.2	1.0	0.4
Std.sapma	0.3	0.29	0.13	0.0	0.1	2.5	0.0	1.1
RMSE (Model)=0.01			Std.sapma=0.13				Ayırma indeksi=10.59	
Güvenirlik=0.99								
Tamamı aynı X^2=496.4		sd=3	p=0.00					
Rasgele normal X^2=3.0		sd=32	p=0.22					

Ayırma indeksi güvenirliği de 0.99 olarak elde edilmiştir. Ayrıca, sabit etki (fixed effect) hipotezi X^2 testi ile reddedilmiştir (X^2=496.4, serbestlik derecesi= 3 ve p=0.00< α=0.05). Buna göre de, puanlayıcıların cömertlik/katılık açısından istatistiksel olarak manidar şekilde birbirlerinden farklı oldukları söylenebilir. Başka bir ifadeyle, matematik performansının ölçülmesinde bulunan dört puanlayıcının puanlaması katılık/cömertlik açısından birbirlerinden farklıdır. Bununla birlikte, puanlayıcıların katılık/cömertlik değerleri 0.5 gibi küçük bir logit aralığında değiştiği için puanlayıcılar arası farklılığın kabul edilebilir ölçüde olduğu söylenebilir (Lee ve Kantor, 2003). Ayrıca, uyum istatistiklerinin işaret ettiği şekliyle; bir bütün olarak puanlayıcıların tutarlı puanlama yapmış oldukları gözlenmektedir. Rasgele etki (random effect) hipotezi X^2 testi ile red edilememiştir (X^2=3.0, serbestlik derecesi= 32 ve p=0.22 > α=0.05). Bu sonuca göre, puanlayıcıların puanlarının normal dağılım gösterdiği ve verilerin modele uyumlu olduğu yorumu yapılabilir. Hem iç uyum (infit) (ortalama=1.0 ve standart sapma=0.1) hem de dış uyum (outfit) (ortalama= 1.0 ve standart sapma= 0.0) değerleri, verinin genel anlamda modele çok iyi uyum gösterdiğine işaret etmektedir. Tablo 14'te de görüleceği gibi, iç uyum ve dış uyum kareler ortalamaları 0.5 ile 1.5 arasında değerler almıştır. Bu değerler, verilerin ölçüm için uygun (verimli) olduğunu işaret etmektedir (Linacre, 2007). Diğer bir deyişle, dört puanlayıcının birbirlerine göre puanlamaları farklılık göstermesine karşın, Rasch modeline göre puanlayıcılar kendi içinde öğrencilerin tümünü, tüm maddeler açısından birbirleriyle tutarlı olarak ölçmüştür yorumu yapılabilir (Nakamura, 2000).

3.3.5. Madde Güçlükleri ve Uygunluk İstatistikleri

Tablo 15'te matematik performansının ölçülmesinin çok değişkenlik kaynaklı Rasch ölçme modelinin analizinde yer alan 18 maddeye ilişkin değerler ve istatistikler yer almaktadır. Tablo 15'te, matematik performansının ölçülmesine ilişkin 18 maddenin ölçme sonuçları verilmiştir. Maddelerin grafiksel gösterimi de Şekil 6'da görülmektedir. Maddelerin güçlük düzeyleri -0.46 ile 0.48 arasında değişmektedir (1 logit birimlik bir değişim gözlenmektedir). 0.48 logit değeri ile yedinci madde (zaman hesaplamalarıyla ilgili bir soru) en zor madde iken, -0.46 logit değeri ile dördüncü madde(geometri sorusu: karenin kenar uzunlukları ve alanı ile ilgili bir soru) en kolay maddedir. Maddelerin logit değerlerinin ortalaması 0.0 ve standart sapması 0.22'dir. Maddelere ilişkin ayırma indeksi 8.36 ve güvenirlik 0.99 olarak elde edilmiştir. Sabit etki (fixed effect) hipotezi X^2 testi ile reddedilmiştir (X^2=1213.3, serbestlik derecesi= 17 ve p=0.00< α=0.05). Buna göre, maddelerin güçlük dereceleri açısından istatistiksel olarak manidar şekile birbirlerinden farklı oldukları söylenebilir. Başka bir ifadeyle, matematik performansının ölçülmesinde yer alan 18 madde güçlük dereceleri açısından birbirlerinden farklıdır. Madde değişkenlik boyutu içinde ayırma indeksinin 8.36 olması, maddelerin güçlük düzeyleri açısından kestirilen hatanın (0.03) sekiz katı büyüklükte bir değişim (varyasyon) olduğunu göstermektedir. Bu sonuç da test geliştiricileri açısından istenen bir durumu işaret etmektedir (Hetherman, 2004).

Rasgele etki (random effect) hipotezi X^2 testi ile red edilememiştir (X^2=17.0, serbestlik derecesi= 16 ve p=0.39 > α=0.05). Bu sonuca göre, maddelerin normal dağılım gösterdiği ve verilerin modelle uyumlu olduğu yorumu yapılabilir. Hem iç uyum (infit) (ortalama=1.0 ve standart sapma=0.2) hem de dış uyum (outfit) (ortalama= 1.0 ve standart sapma= 0.2) değerleri, verinin genel anlamda modele çok iyi uyum gösterdiğine işaret etmektedir. Ayrıca, maddelere ilişkin iç uyum ve dış uyum kareler ortalamaları 0.5 ile 1.5 arasında değerler almıştır. Bu değerler, verilerin ölçüm için uygun (verimli) olduğunu işaret etmektedir (Linacre, 2007).

Tablo 15. Matematik Performansının Ölçülmesinde Yer Alan 18 Maddenin Ölçme Raporu

			Madde Katılığı		Uygunluk İçi		Uygunluk Dışı	
Madde No	Madde Ort.	Madde Toplamr	Logit Ölçüsü	S.H.	Kareler Ort.	Z Std.	Kareler Ort.	Z Std.
7	2.6	2.65	0.48	0.02	1.0	0	1.0	0
15	2.9	3.18	0.34	0.02	0.8	-5	0.6	-5
9	3.3	3.75	0.16	0.02	1.0	0	1.0	0
12	3.4	3.82	0.14	0.02	0.8	-4	0.9	-1
1	3.4	3.86	0.12	0.02	1.0	0	1.2	2
17	3.5	3.94	0.09	0.02	0.9	-2	0.9	-1
6	3.6	4.06	0.04	0.02	0.9	-2	0.8	-1
8	3.7	4.13	0.00	0.02	1.1	1	1.0	0
14	3.7	4.14	0.00	0.02	0.9	-1	0.8	-1
18	3.7	4.14	0.00	0.02	1.0	0	1.1	0
2	3.7	4.17	-0.02	0.03	0.8	-3	0.9	-1
5	3.8	4.22	-0.04	0.03	1.3	5	1.2	1
16	3.7	4.21	-0.04	0.03	1.1	2	1.2	1
3	3.9	4.33	-0.11	0.03	1.2	3	1.3	2
11	3.9	4.33	-0.11	0.03	1.1	1	1.1	1
13	4.2	4.54	-0.29	0.03	1.1	1	0.9	-1
10	4.2	4.56	-0.31	0.03	1.2	3	1.0	0
4	4.4	4.67	-0.46	0.03	1.1	0	1.4	2
Ortalama	3.6	4.04	0.0	0.03	1.0	0.0	1.0	-0.1
Std.sapma	0.4	0.47	0.22	0.0	0.2	2.9	0.2	2.1
RMSE (Model)=0.03			Std.sapma=0.21				Ayırma indeksi=8.36	
Güvenirlik=0.99								
Tamamı aynı X^2=1213.3		sd=17	p=0.00					
Rasgele normal X^2=17.0		sd=16	p=0.39					

3.4. ALT PROBLEM IV.: KLASİK TEST KURAMI, GENELLENEBİLİRLİK KURAMI VE ÇOK DEĞİŞKENLİK KAYNAKLI RASCH ÖLÇME MODELİNE GÖRE BİREY (ÖĞRENCİ) VE PUANLAYICI DEĞİŞKENLİK BOYUTLARINDA ELDE EDİLEN PARAMETRELERİN KARŞILAŞTIRILMASI

2007 yılında uygulanan matematik performansının ölçülmesine ilişkin olarak yapılan klasik test kuramı, genellenebilirlik kuramı ve çok değişkenlik kaynaklı Rasch ölçme modeline bağlı yapılan analizler sonucunda elde edilen istatistikler öğrenci ve puanlayıcı değişkenlik boyutlarına göre sırasıyla aşağıda karşılaştırılmıştır.

3.4.1. Birey (Öğrenci) Boyutu

Klasik test kuramında, öğrencilere ilişkin elde edilen gözlenen puanlar varyansının gerçek puanlar varyansına yaklaşma olasılığı, gözlenen puanların elde edildiği testin güvenirliği arttıkça artmaktadır (Baykul, 2000). Bir başka deyişle, ölçme aracının güvenirliği azaldıkça, ölçme aracıyla ölçülen özellik açısından bireyler arası farklılığın görülebilme olasılığı azalmaktadır. Burdadan da anlaşılacağı üzere, klasik test kuramında bireylerin gözlenen puanları ölçme aracına bağımlıdır. Ölçme aracının güvenirliği ölçüsünde, bireylere ilişkin güvenilir sonuçlar elde edilebilmektedir. Klasik kuramda, ölçme aracının iç tutarlılığı anlamında hesaplanan Cronbach alfa güvenirlik katsayısı, aynı zamanda ölçülen özellik açısından bireylere ilişkin ne kadar güvenilir bilgi elde edilebildiğinin de bir ölçüsü olmaktadır.

Performansın ölçülmesinde kaçınılmaz olan bir başka değişkenlik kaynağı da puanlayıcılardır. Ancak klasik test kuramında, bireyler, maddeler ve puanlayıcıların aynı anda göz önünde bulundurularak hesaplanabilen bir güvenirlik katsayısı bulunmamaktadır. Farklı değişkenlik kaynakları için farklı güvenirlik katsayılarının hesaplanması gerekmektedir. Klasik test kuramında, performansın ölçülmesine ilişkin iç tutarlılığa bakılacaksa ve farklı farklı puanlayıcıların puanlaması söz konusu ise, her bir puanlayıcı için ayrı bir iç tutarlılık katsayısı hesaplamak kaçınılmazdır. Buna göre, bu çalışmada yer alan matematik performansının ölçülmesinde yer alan 18 maddeye ilişkin iç tutarlılık anlamında Cronbach alfa katsayısı birinci ve ikinci puanlayıcılar için 0.91, üçüncü ve dördüncü puanlayıcılar için de 0.92 olarak hesaplanmıştır. Bu değerlere göre, 2007 yılında yapılan matematik perfromansının ölçülmesiyle, matematik performansına ilişkin elde edilen öğrenci puanlarının güvenilir olduğu yorumu yapılabilir.

Genellenebilirlik analizine dayalı olarak yapılan G çalışması sonucunda matematik performansının ölçülmesinin tek boyutu için kestirilen varyans ve toplam varyansı açıklama oranları incelendiğinde, birey (b) ana etkisi için kestirilen varyans bileşeninin (1.160) toplam varyansın % 32.6'sını açıkladığı görülmektedir. G çalışmasıyla bireyler için kestirilen varyans bileşenini, toplam varyans içinde en yüksek ikinci sırada paya sahip varyans bileşenidir. Genellenebilirlik çalışmalarında, birey ana etkisi evren puanı varyansı olarak değerlendirilir ve ölçülen özellik açısından bireyler arası farklılaşmayı

ifade eder (Shavelson ve Webb, 1991; Brennan 2001). Bireyler için kestirilen varyansın toplam varyans içindeki oranının en büyük olması istenilen bir durumdur. Bu durum, ölçme ile elde edilen boyutta bireyler arası farklılıkların ortaya çıkarılabildiğinin bir göstergesidir.

Genellenebilirlik kuramında G-katsayısı, klasik test kuramındaki, gerçek varyansın gözlenen varyansa oranı olan, güvenirlik katsayısıyla benzerlik gösterir. Özellikle, (1) değişkenlik kaynağının maddeler olduğu, (2) D-çalışmasındaki düzeylerin testte yer alan maddelerin sayısına eşit olduğu ve (3) sadece göreceli modelin kullanıldığı durumda, tek değişkenlik kaynaklı (maddeler) çaprazlanmış desenlerdeki G-katsayısı, klasik test kuramındaki Cronbach alfa'nın eşiti olmaktadır (Sudweeks, Reeve ve Bradshaw, 2005). 2007 yılında yapılan matematik perfromansının ölçülmesiyle, matematik performansına ilişkin 18 madde ve dört puanlayıcı üzerinden elde edilen G-katsayısı 0.92 olarak bulunmuştur. Bu sonuca göre 2007 yılında yapılan matematik perfromansının ölçülmesiyle, öğrencilerin matematik performansına ilişkin güvenilir sonuçlar elde edildiği yorumu yapılabilir.

Çok değişkenlik kaynaklı Rasch ölçme modeli analizlerine göre, 2007 yılında yapılan matematik performansının ölçülmesine ilişkin 18 madde ve dört puanlayıcı üzerinden elde edilen ayırma indeksi (G) 4.39 ve güvenirliği 0.95 olarak yeterince yüksek bulunmuştur. Ayırma indeksi, değişkenlik kaynaklarına ilişkin elde edilen ölçmelerin sürekliliği boyunca ne ölçüde birbirlerinden ayrıldıklarını gösteren bir istatistiktir (Turner, 2003). Öğrenci değişkenlik boyutu için ayırma indeksi; KR-20, Cronbach alfa ve genellenebilirlik katsayısına eşit olarak yorumlanabildiği için, testin iç tutarlılık anlamında güvenirliliğin yüksek olduğu yorumu yapılabilir (Nakamura, 2000).

Tablo 16. Üç Kurama İlişkin Analizlere Göre Öğrenci Boyutuna İlişkin İstatistiklerin Özeti

Klasik Test Kuramı	Genellenebilirlik Kuramı	ÇDKRÖM
Cronbach alfa güvenirlik katsayısı: 0.92	G katsayısı: 0.92	Öğrenci değişkenlik boyutu için ayırma indeksi güvenirliği: 0.95

3.4.2. Puanlayıcı Boyutu

Klasik test kuramında, ikiden fazla puanlayıcının belirli bir davranışı, performansı, soruyu vb. puanlaması durumunda, puanlayıcılar arası güvenirliğin belirlenmesinde parametrik olmayan bir istatistiksel teknik olan Kendall'ın uyuşum katsayısı hesaplanabilir (Howell, 2002). Ancak, performansın ölçülmesinde puanlayıcılar önemli bir hata kaynağı olmasına karşın aynı zamanda farklı hata kaynakları ve bunlar arasındaki etkileşimden doğan hata kaynakları da bulunmaktadır. Fakat Kendall'ın uyum katsayısı ile güvenirliğin hesaplanmasında, birden fazla hata kaynağı ve bunlar arasındaki etkileşim birlikte göz önünde bulundurulmamaktadır. Böylece klasik test kuramında, farklı varyans kaynaklarının büyüklüğünün kestirilmesi çoklu analizlerin yapılmasını gerektirdiği gibi, farklı varyans kaynakları arasındaki etkileşim de kestirilememektedir. Klasik test kuramı analizlerine göre, 2007 yılında yapılan matematik performansının ölçülmesine ilişkin elde edilen puanlayıcılar arası güvenirliğin bir ölçüsü olarak hesaplanan Kendall'ın uyum katsayısı 0.52 olarak bulunmuştur (X^2=315.16, sd=3, p=0.00<0.05). Buna göre puanlayıcılar arası uyum olduğu yorumu yapılabilir (Howell, 2002). Ayrıca, çalışmada yer alan dört puanlayıcının verdikleri bunlar arasındaki korelasyon değerlerine bakıldığında en küçük değerin 0.90 (birinci ve ikinci puanlayıcılara arasında) ve en yüksek değerin de 0.97 (ikinci ve dördüncü puanlayıcılara arasında) olduğu görülmektedir. Bu değerlerden de anlaşılacağı üzere, dört puanlayıcının verdikleri puanlar arasındaki ilişki neredeyse 1'e yakın ve yüksek değerlere sahiptir. Bu da puanlayıcıların verdikleri puanlar arasında büyük bir paralellik olduğuna işaret etmektedir. Bir başka deyişle bir puanlayıcının verdiği puanların sıralaması ile diğer bir puanlayıcının verdiği puanların sıralaması paralellik göstermektedir. Fakat buradan, puanlayıcıların verdikleri puanlar arasında bir farklılık olmadığına dair bir yorum yapılamaz. Her bir puanlayıcının verdiği puanların ortalaması arasında fark olup olmadığı testi ilişkili örneklem üzerinde tek değişkenli varyans analizi ile test edilmiş, yokluk hipotezi rededilerek, istatistiksel olarak anlamlı bir fark olduğu sonucuna varılmıştır (F=13.801, p=0.00< α=0.05). Sonuç olarak, matematik performansının ölçülmesinde yer alan dört puanlayıcının verdikleri puanların sıralaması paralellik göstermekle birlikte, puanlamadaki katılık/cömertlik düzeyleri birbirlerinden farklıdır.

Genellenebilirlik kuramında bir öğrencinin, performansının ölçülmesinde yer alan bir görevde gösterdiği performansı, puanlaycı ve görev gibi olası tüm değişkenlik kaynaklarının bir arada bulunduğu karmaşık bir evrendenki performansından çekilen bir örneklemi olarak görülür. Böylece tek bir analizle tüm değişkenlik kaynaklarına ilişkin ayrı ayrı ve bu değişkenlik kaynaklarının birbirleriyle etkileşimine ilişkin varyansların hesaplanmasına ve tüm değişkenlik kaynaklarını bir arada göz önünde bulundurarak tek bir güvenirlik katsayısının hesaplanmasına imkan tanır. Genellenebilirlik kuramı analizlerine göre, 2007 yılında yapılan matematik performansının ölçülmesine ilişkin elde edilen puanlayıcı ana etkisi için kestirilen varyans bileşeni (0.075) toplam varyansın % 2.1'ini açıklayarak, ana etkiler içinde en düşük varyans bileşeni olmaktadır. Puanlayıcı ana etkisinin G çalışması ile kestirilen varyans oranının düşük çıkması, puanlayıcılar arası değişkenliğin çok az olduğu, puanlayıcıların verdikleri puanlar arasında bir tutarlılığın söz konusu olduğu şeklinde yorumlanabilir.

Çok değişkenlik kaynaklı Rasch ölçme modeli her bir değişkenlik kaynağının kendi içinde olduğu kadar, değişkenlik kaynaklarının her birine ilişkin olarak da örneklemden bağımsız parametre kestiriminde bulunmayı sağlar. Değişkenlik kaynaklarının her biri aynı zamanda analiz edilir ve bu analizler istatistiksel olarak birbirinden bağımsızdır. Bir başka deyişle, genellenebilirlik kuramı değişkenlik kaynakları arasındaki etkileşimi de göz önünde bulunduran bir yaklaşımken; madde tepki kuramı, etkileşimin olmadığı varsayımına dayalı bir yaklaşımdır. Tüm değişkenlik kaynakları tek bir genel doğrusal ölçek üzerinde yer alır. Bireylerin yetenekleri, belirli maddelerin dağılımlarının özelliklerinden, belirli puanlayıcıların performansa verdikleri puanlardan vb. serbest olarak kestirilir. Benzer şekilde, maddelerin güçlük düzeylerinin ve puanlayıcılarının sertlik düzeylerinin kestirimi de verilerin elde edildiği desende yer alan diğer değişkenlik kaynaklarının dağılım özelliklerinden serbest olarak yapılmaktadır (Smith ve Kulikowich, 2004).

Çok değişkenlik kaynaklı Rasch ölçme modeli analizlerine göre, 2007 yılında yapılan matematik performansının ölçülmesine ilişkin puanlayıcı boyutunda elde edilen ayırma indeksi 10.59 gibi oldukça yüksek bir değer olarak elde edilmiştir. Bu değer puanlayıcılar arası farklılığın oldukça fazla olduğunun bir göstergesidir (Nakamura, 2002). Ayırma indeksi güvenirliği ise 0.99 olarak bulunmuştur. Ayrıca, sabit etki (fixed

effect) hipotezi X^2 testi ile reddedilmiştir (X^2=496.4, serbestlik derecesi= 3 ve p=0.00<α=0.05). Buna göre, puanlayıcıların cömertlik/katılık açısından istatistiksel olarak manidar şekile birbirlerinden farklı oldukları söylenebilir. Başka bir ifadeyle, matematik performansının değerlendirmesinde bulunan dört puanlayıcının puanlaması katılık/cömertlik açısından birbirlerinden farklıdır. Hem iç uyum (infit) (ortalama=1.0 ve standart sapma=0.1) hem de dış uyum (outfit) (ortalama= 1.0 ve standart sapma= 0.0) değerleri, verinin genel anlamda modele çok iyi uyum gösterdiğine işaret etmektedir. Bu sonuçlara göre, Rasch modeli analizine dayalı olarak dört puanlayıcının, öğrencileri tutarlı olarak ölçtüğü yorumu yapılabilir. Elde edilen bu bulgulara göre, matematik performansının ölçülmesinde puanlama yapan dört puanlaycının puanlaması katılık/cömertlik açısından farklılık göstermekle birlikte, öğrencilere verdikleri puanların birbiriyle tutarlı olduğu sonucuna varılabilir.

Tablo 17. Üç Kurama İlişkin Analizlere Göre Puanlayıcı Boyutuna İlişkin İstatistiklerin Özeti

Klasik Test Kuramı	Genellenebilirlik Kuramı	ÇDKRÖM
Kendall'ın uyum katsaysı: 0.518 Puanlayıcılar arası korelasyon katsayıları:0.90-0.97	Puanlayıcı değ.kay.varyansı: 0.075 (%2.1)	Puanlayıcı değişkenlik boyutu için ayırma indeksi güvenirliği: 0.99

BÖLÜM IV

SONUÇ VE ÖNERİLER

Bu bölümde, araştırmanın bulguları üzerinden elde edilen sonuçlar özetlenmiş ve bu sonuçlara dayalı yapılabilecek önerilere yer verilmiştir.

4.1. SONUÇLAR

Araştırmadan elde edilen bulgulara dayalı olarak ulaşılan sonuçlara alt problemlerin sırasına uygun olarak aşağıda açıklanmıştır.

I. 1) 2007 yılında yapılan matematik performansının ölçülmesinde yer alan maddelerin öğrencilerin matematik başarısını, var olan boyutta ölçüp ölçmediği hem açıklayıcı hem de doğrulayıcı faktör analizi çalışması ile incelenmiştir. Matematik performansının ölçülmesinde yer alan 18 maddeye ilişkin öncelikle yapılan açıklayıcı faktör analizi sonuçlarına göre maddelerin tümü "matematik başarısı" adı altında tek bir boyutta yer almıştır. Açıklayıcı faktör analizi ile elde edilen bu tek boyut, doğrulayıcı faktör analizi ile test edilmiştir. Bu çalışmanın sonucu da matematik performansının ölçülmesinde yer alan 18 maddenin tek boyutta toplandığını doğrular nitelikte olduğu görülmüştür.

I. 2) 2007 yılında uygulanan matematik performansının ölçülmesinin, klasik test kuramına göre iç tutarlılık düzeyinin belirlenmesi için Cronbach α güvenirlik katsayısı hesaplanmıştır. Tüm puanlayıcılar için ayrı ayrı ve puanlayıcıların her bir maddeye verdikleri puanların ortalamalarından elde edilen güvenirlik katsayılarına göre, matematik performansının ölçülmesinde yer alan maddelerin matematik başarısını oldukça tutarlı bir şekilde ölçtüğü sonucuna varılmıştır.

I. 3) 2007 yılında yapılan matematik performansının ölçülmesinde yer alan dört puanlayıcı arasındaki tutarlılığın belirlenmesi için klasik test kuramına göre öncelikle parametrik olmayan istatistiksel bir teknik olan Kendall'ın uyum katsayısı kullanılmıştır. Ayrıca, her bir puanlayıcının verdiği puanlar ile diğer bir puanlayıcının verdiği puanlar arasındaki korelasyon katsayıları hesaplanmış ve her bir puanlayıcının verdiği puanların ortalaması arasındaki farklılığın olup olmadığı testi ilişkili örneklem

üzerinde tek değişkenli varyans analizi ile test edilmiştir. Yapılan bu çalışmalara göre, matematik performansının ölçülmesinde yer alan dört puanlayıcının puanlamadaki katılık/cömertlik düzeylerinin birbirinden farklı olduğu görülmekle birlikte puanlayıcıların verdikleri puanların sıralaması paralellik göstermekte olup, birbirleriyle tutarlı olacak şekilde puanlama yaptıkları sonucuna varılmıştır.

II.1) 2007 yılında uygulanan matematik performansının ölçülmesinde yer alan 18 maddeye tümüyle çaprazlanmış b x g x p modeli uygulanarak yapılan G çalışması sonucunda matematik performansının ölçülmesinin tek boyutu için kestirilen varyans ve toplam varyansı açıklama oranları incelendiğinde, birey (b) ana etkisi için kestirilen varyans bileşeninin toplam varyansın büyük bir oranını açıkladığı görülmüştür. Tek değişkenli modelle bireyler için kestirilen varyans bileşeni, toplam varyans içinde en yüksek ikinci sırada paya sahip varyans bileşeni olmuştur. Bireyler için kestirilen varyansın toplam varyans içindeki oranının en büyük olması istenilen bir durumdur. Buna göre, yapılan matematik performansının ölçülmesinin, elde edilen boyutta bireyler arası farklılıkları ortaya çıkarılabildiği sonucuna varılmıştır. Görev ana etkisi için kestirilen varyans, toplam varyansın küçük bir oranını açıklamaktadır. Buna göre, matematik performansında yer alan görevlerin güçlük düzeylerinin birbirine yakın olduğu, bir görevin diğerinden daha zor olmadığı sonucuna varılmıştır. Puanlayıcı ana etkisi için kestirilen varyans bileşeninin toplam varyansın çok küçük bir oranını (sıfıra çok yakın) açıkladığı görülmüştür. Buradan, puanlayıcıların tüm öğrenciler için yaptıkları puanlamalar arasında farklılığın hemen hemen hiç olmadığı, puanlamalar arasında bir tutarlılığın söz konusu olduğu sonucu çıkarılmıştır.

II.2) Matematik performansının ölçülmesinde tek boyut altında yer alan 18 madde sayısına ve bu maddeleri puanlayan 4 puanlayıcıya göre elde edilen mutlak hata ve bağıl hata varyansları birbirlerine çok yakın ve oldukça düşük değerler olarak elde edilmiştir. Puanlayıcı varyansının toplam varyansa oranının sıfıra oldukça yakın olması sebebiyle, puanlayıcı sayısındaki artış ya da azalma mutlak ve bağıl hata varyansını fazlaca etkilememekte, madde (görev) varyansının toplam varyansa oranı puanlayıcı varyansına göre daha büyük olduğundan, madde sayısındaki artış ya da azalma mutlak ve bağıl hata varyansını daha fazla etkilemektedir. Buna göre, mutlak ya da bağıl hata varyansının azaltılmasının istenildiği bir durumda sadece iki puanlayıcı kullanılarak ve madde

sayısını üçte iki arttırarak (mutlak hata varyansı:0.127 ve bağıl hata varyansı: 0.084) var olan durumdaki (18 madde ve dört puanlayıcı) hata varyanslarının (sırasıyla; 0.133 ve 0.105) altında hata varyansına ulaşılabileceği sonucuna varılmıştır.

Yukarıdaki sonuca paralel olarak, matematik performansının ölçülmesinde tek boyut altında yer alan 18 madde sayısına ve bu maddeleri puanlayayan 4 puanlayıcıya göre elde edilen genellenebilirlik (G katsayısı) ve güvenirlik (Φ katsayısı) birbirlerine çok yakın ve oldukça yüksek değerler olarak elde edilmiştir. Puanlayıcı varyansının toplam varyansa oranının sıfıra oldukça yakın olması sebebiyle, puanlayıcı sayısındaki artış ya da azalma genellenebilirlik ve güvenirlik katsayılarını fazlaca etkilememekte, madde (görev) varyansının toplam varyansa oranı puanlayıcı varyansına göre daha büyük olduğundan, madde sayısındaki artış ya da azalma genellenebilirlik ve güvenirlik katsayılarını daha fazla etkilemektedir. Buna göre, genellenebilirlik ve güvenirlik katsayılarının arttırılmasının istenildiği bir durumda sadece iki puanlayıcı kullanılarak ve madde sayısını üçte iki arttırarak (G katsayısı: 0.932 ve Φ katsayısı: 0.8901) var olan durumdaki (18 madde ve dört puanlayıcı) genellenebilirlik ve güvenirlikten (sırasıyla; 0.917 ve 0.897) daha yüksek değerlere ulaşılabileceği belirlenmiştir. Matematik performansının ölçülmesinde daha yüksek genellenebilirlik ve güvenirliğe ulaşmak için puanlayıcı sayısını arttırmak yerine madde sayısını arttırmanın daha ekonomik ve kullanışlı olacağı sonucuna varılmıştır.

III. 1) Çok değişkenlik kaynaklı Rasch ölçme modeline göre, 2007 yılında uygulanan matematik performansının ölçülmesinin FACET analizi ile model veri uyumununa bakılmış ve modelin veri için uygun olduğu sonucuna varılmıştır.

III. 2) Matematik performansının ölçülmesinin uygulandığı toplam 203 öğrencinin çok değişkenlik kaynaklı Rasch ölçme modeline göre, FACET analizi ile ayırma indeksi hesaplanmış ve bu değerin oldukça yüksek olduğu belirlenmiştir. Ayırma indeksi, değişkenlik kaynaklarına ilişkin elde edilen ölçmelerin sürekliliği boyunca ne ölçüde birbirlerinden ayrıldıklarını veren bir istatistik olup, öğrenci değişkenlik boyutu için ayırma indeksi güvenirliği; KR-20, Cronbach alfa ve genellenebilirlik katsayısına benzer yorumlanmaktadır. Buna göre, matematik performansının ölçülmesinin çok

değişkenlik kaynaklı Rasch ölçme modeline göre iç tutarlılık anlamında güvenirliğinin yüksek olduğu sonucuna varılmıştır.

III. 3) Matematik performansının ölçülmesinde yer alan dört puanlayıcının, çok değişkenlik kaynaklı Rasch ölçme modeline göre FACETS analizi ile hesaplanan ayırma indeksinin oldukça yüksek bir değerde olduğu belirlenmiştir. Bu değere göre puanlayıcıların cömertlik/katılık açısından birbirlerinden farklı oldukları sonucuna varılmıştır. Bununla birlikte, uyum istatistiklerinin işaret ettiği şekliyle; bir bütün olarak puanlayıcıların tutarlı puanlama yapmış oldukları belirlenmiştir. Diğer bir deyişle, dört puanlayıcının birbirlerine göre puanlamaları farklılık göstermesine karşın, Rasch modeline göre her bir puanlayıcı kendi içinde öğrencilerin tümünü, tüm maddeler açısından tutarlı olarak ölçtüğü sonucu elde edilmiştir.

III. 4) Matematik performansının ölçülmesinin çok değişkenlik kaynaklı Rasch ölçme modeline göre FACETS analiziyle maddelerin güçlük dereceleri açısından birbirlerinden farklı oldukları sonucuna varılmıştır. Başka bir ifadeyle, matematik performansının ölçülmesinde yer alan 18 madde güçlük dereceleri açısından birbirlerinden farklıdır sonucuna varılmıştır.

IV) 2007 yılında uygulanan matematik performansının ölçülmesine ilişkin olarak yapılan klasik test kuramı, genellenebilirlik kuramı ve çok değişkenlik kaynaklı Rasch ölçme modeline bağlı yapılan analizler sonucunda elde edilen istatistikler öğrenci ve puanlayıcı değişkenlik boyutlarına göre karşılaştırılmıştır.

Öğrenci boyutunda, tüm kuramların karşılaştırılabilir değerleri olarak, klasik test kuramına göre Cronbach alfa katsayısı 0.92, genellenebilirlik kuramına göre genellenebilirlik katsayısı 0.92 ve çok değişkenlik kaynaklı Rasch ölçme modeline göre öğrenci değişkenlik kaynağına ilişkin ayırma indeksi güvenirliği 0.95 olarak bulunmuştur. Elde edilen bu değerlere göre ölçme sonuçlarının iç tutarlılık anlamında güvenirliğinin oldukça yüksek olduğu sonucuna varılmıştır.

Puanlayıcı boyutunda, tüm kuramların karşılaştırılabilir olan değerleri birlikte göz önüne alınmış ve şu sonuçlara varılmıştır: klasik test kuramına göre Kendall'ın uyum katsayısı 0.52 olarak hesaplanmıştır. Puanlayıcılar arası korelasyon katsayılarının 0.90

ile 0.97 arasında değiştiği gözlenmiş ve puanlayıcıların verdikleri puanların ortalamaları arasındaki farlılık için uygulanan F testi sonucunda istatistiksel olarak anlamlı bir fark olduğu sonucuna varılmıştır. Tüm bu elde edilen sonuçlara dayalı olarak; dört puanlayıcının verdikleri puanların ortalamaları arasında farklılık olduğu ancak puanlayıcıların birbirleriyle tutarlı olacak şekilde puanlama yaptıkları gözlenmiştir.

Genellenebilirlik kuramına göre, puanlayıcı değişkenlik kaynağına ilişkin kestirilen varyansın %2 gibi çok küçük bir değer olması ve tüm puanlara ilişkin genellenebilirlik katsayısının oldukça yüksek bir değer çıkması sonucunda, puanlar arasında puanlayıcılardan kaynaklı bir değişimin gözlenmediği sonucuna varılmıştır.

Çok değişkenlik kaynaklı Rasch ölçme modeline göre, puanlayıcı değişkenlik kaynağına ilişkin ayırma indeksi güvenirliği 0.99 olarak hesaplanmıştır. Bu değer puanlayıcıların verdikleri puanlar arasında farklılık olduğunu göstermiştir. Ancak puanlayıcı değişkenlşk kaynağına ilişkin uyum istatistiklerinin 0.5 ile 1.5 arasında değerler almış olması puanlayıcıların verdikleri puanlarda katılık/cömertlik açısından birbirleriyle tutarlı puanlama yaptıklarına işaret etmektedir.

4.2. ÖNERİLER

A. Matematik Performansının Ölçülmesine İlişkin Öneriler

1. Matematik performansının ölçülmesinde zaman, ekonomi ve emek açısından dört puanlayıcı kullanmak yerine daha az sayıda puanlayıcı kullanarak ve madde sayısını arttırarak istenilen güvenirlikte ölçme yapmak daha uygun olabilir.

2. Matematik performansının ölçülmesinde puanlayıcıların, puanlama yöntemine ilişkin açık, anlaşılır ve yeterli derecede bilgilendirilmesi, puanlayıcılar arası uyumun istenilen düzeye ulaşabilmesi açısından yararlı olabilir.

İleride Yapılabilecek Çalışmalar İçin Öneriler:

1. Matematik performansının puanlanmasında kullanılan bütünsel dereceli puanlama anahtarının yanısıra analitik dereceli puanlama anahtarıyla puanlama yapılarak benzer bir çalışma uygulanabilir.

2. Matematik performansının puanlanmasında hem bütünsel hem de analitik dereceli puanlama anahtarı kullanılarak aynı çalışma üzerinde bu iki puanlama yöntemi karşılaştırılarak benzer bir çalışma uygulanabilir.

3. Matematik dersi dışında farklı derslerde uygulanan performans ölçmeleri için benzer güvenirlik çalışmaları yapılabilir.

4. Performansın ölçülmesine ilişkin yapılan uygulamalarında öğrencilerin güdülenmişlik düzeyini arttırmak böylece elde edilecek puanların güvenirliğinin artmasını sağlamak için öğrencilere ödül vb... sunularak benzer çalışmalar yapılabilir.

B. Klasik Test Kuramı, Genellenebilirlik Kuramı ve Çok Değişkenlik Kaynaklı Rasch Ölçme Modelinin Kullanımına İlişkin Öneriler

1. Puanlayıcılar arası tutarlılığın araştırılması söz konusu olduğunda genellenebilirlik kuramı ile çok değişkenlik kaynaklı Rasch ölçme modelinden hangisinin kullanılmasının uygun olacağını belirlemede bazı durumlar yol gösterici olabilir. Var olan çalışmada yer alan puanlayıcıların verdikleri puanlar arasında bir uyum söz konusuysa ve elde edilen puanlarla evrendeki tüm puanlayıcılara genelleme yapılmak isteniyorsa, genellenebilirlik kuramından yararlanmak uygun olabilir. Veriler modele uyumlu olduğunda, çok değişkenlik kaynaklı Rasch ölçme modeli kullanılarak, diğer değişkenlik kaynaklarıyla birlikte (bireyler, maddeler vb.) aynı doğrusal ölçekte yer alan puanlayıcı katılık/cömertlik kestirimlerine ilişkin yorumlar yapılabilir.

2. Performansın ölçülmesile elde edilen puanların hangi ölçek yapısına uygun olduğu, ölçmenin hangi amaçla yapıldığı ve elde edilen verilerin kullanılacak kuramın varsayımlarını sağlayıp sağlamadığı göz önüne alınarak analiz için hangi kuramın kullanılmasının uygun olacağına karar verilebilir.

3. Yapılan performans ölçme sonuçlarına göre hem bağıl hem de mutlak kararlara varmak söz konusu ise genellenebilirlik kuramının kullanılması tercih edilebilir.

4. Geleneksel anlamdaki güvenirlik ve geçerlik arasındaki farklılığı ortadan kaldırdığı için genellenebilirlik kuramı, tek bir çalışmayla güvenirliğin ve geçerliğin belirlenmesini sağlayarak daha ekonomik ve pratik bir kuram olarak önerilebilir.

5. Genellenebilirlik kuramında yer alan farklı karar çalışmalarıyla farklı sayıdaki madde (görev) ya da puanlayıcı sayılarına dayalı güvenirlik tahminleri yapılarak sonraki ölçme çalışmaları için en uygun ve ekonomik puanlayıcı ve madde (görev) sayılarına önceden karar verilebilir.

6. Çok değişkenlik kaynaklı Rasch ölçme modeli, tek bir çalışma ile her bir değişkenlik kaynağına ilişkin farklı güvenirliklerin hesaplanmasına imkan verdiği için her bir değişkenlik kaynağı için ayrıntılı bilgi sağlamaya olanak tanır. Tüm değişkenlik kaynaklarının tek bir ölçekte birlikte değerlendirilmesi, her bir değişkenlik kaynağı içindeki uyumsuzlukların belirlenmesi ve böylece hangi madde ya da puanlayıcıdan kaynaklı bir sorun olduğunun açıkça görülmesinin istendiği durumlarda çok değişkenlik kaynaklı Rasch ölçme modelinin kullanılması uygun olabilir.

7. Tüm kuramların yararları göz önüne alındığında, yapılacak performans ölçme çalışmalarında sadece klasik test kuramına dayalı güvenirliğin belirlenmesi yerine diğer kuramlardan en az birinin de beraberinde kullanılmasının yararlı olacağı söylenebilir.

İleride Yapılabilecek Çalışmalar İçin Öneriler:

Bu çalışmada matematik performansının ölçülmesinin güvenirliği genellenebilirlik kuramının b x g x p modeli ile analiz edilmiştir. Benzer çalışmalar genellenebilirlik kuramının farklı modelleri ve farklı değişkenlik kaynakları ele alınarak yapılabilir.

KAYNAKÇA

Alharby, E. R. (2006). *A Comparison Between Two Scoring Methods, Holistic vs. Analytic Using Two Measurement Models, The Generalizability Theory and The Many Facet Rasch Measurement Within The Context of Performance Asssessment.* Yayınlanmamış doktora tezi. The Pennsylvenia State University.

Allal, L. ve Cardinet, J. (1997). Generalizability Theory. *Educational Research Methodology and Measurement an International Handbook.* (Second Editon). Editor Keeves, J. P. Cambridge, UK.

Anastasi, A. ve Urbina, S. (1997). *Psychological Testing* (7. Basım). Upper Saddle River, NJ.: Prentice Hall.

Atılgan, H., Kan, A. ve Doğan, N. (2006). *Eğitimde Ölçme ve Değerlendirme.* Ankara: Anı Yayıncılık.

Atılgan, H. (2005). Genellenebilirlik Kuramı ve Puanlayıcılar Arası Güvenirlik için Örnek Bir Uygulama. *Eğitim Bilimleri ve Uygulama*, 4 (7).

Atılgan, H. (2004). *Genellenebilirlik Kuramı ve Çok Değişkenlik Kaynaklı Rasch Modelinin Karşılaştırılmasına İlişkin Bir Araştırma.* Yayınlanmamış doktora tezi. Hacettepe Üniversitesi: Ankara.

Baker, E. L., O'Neil, H. F. ve Linn, R. L. (1993). Policy and Validity Prospects for Performance-Based. *American Psychologist*, 48, 1210-1218.

Baykul, Y., Gelbal, S. ve Kelecioğlu, H. (2003). *Anadolu Lisesi Öğretmenleri İçin Eğitimde Ölçme ve Değerlendirme.* Milli Eğitim Basımevi, İstanbul.

Baykul, Y. (2000). *Eğitimde ve Psikolojide Ölçme: Klasik Test Teorisi ve Uygulaması*. Ankara: ÖSYM.

Bloom, B., Engelhart, M., Furst, E., Hill, W. ve Krathwohl, D. (1956). *Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook I*: Cognitive Domain. New York, Toronto: Longmans, Green.

Boring, R. L. (2002). *Human and Computerized Essay Assessment: A Comparative Analysis of Holistic, Analytic and Latent Semantic Methods.* Yayınlanmamış doktora tezi. New Mexico State University, New Mexico.

Brennan, R. L. (2001). *Generalizability Theory*. New York: Springer-Verlog.

Brown, W. L., O'Gorman, K. ve Du, Y. (1996). The Reliability and Validity of Mathematics Performance Assessment. *Paper for Presentation of The Annual Conference American Educational Research Association*, New York.

Brookhart, S. M. (1999). The Art and Science of Classroom Assessment: The Missing Part of Pedagogy. *ASHE-ERIC Higher Education Report* (Vol.27, 1). Washington DC: The George Washington University, Graduate School of Education and Human Development.

Büyüköztürk, Ş. (2006). *Sosyal Bilimler İçin Veri Analizi El Kitabı, İstatistik, Araştırma Deseni, SPSS Uygulamaları ve Yorumu.* (6. Baskı). PegemA Yayıncılık, Ankara.

Clauser, B. E. (2000). Recurrent Issues and Recent Advances in Scoring Performance Assessment. *Applied Psychological Measurement*, 24, 310-324.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20, 37-46.

Cooper, M. (1997). Nonparametric and Distribution-free Statistics*. Educational Research Methodology and Measurement an International Handbook.* (Second Editon). Editor Keeves, J. P. Cambridge, UK.

Cronbach, L. J. (1980). *Validity on Parole: How Can We Go Straight*. In W. B. Schrader (Editör). New Directions for Testing and Measurement: Measuring Progress Over A Decade. San Francisco: Jassey-Bass.

Cronbach, J. L., Gleser, G. C., Nanda, H. ve Rajaratman, N. (1972*). The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. New York: John Wiley and Sons.

Crocker, L. ve Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. Harcourt Brace Javanovich College Publishers, USA.

Embertson, E. E. ve Steven, P. R. (2000). *Item Response Theory for Psychologists*. Lawrence Erlbaum Associates, Publishers, Mahwah, New Jersey, London.

Engelhard, Jr. ve Myford, M. C. (2003). *Monitoring Faculty Consultant Performance in The Advanced Placement English Literature and Composition Program with A Many-faceted Rasch Model.* (College Board Research Rep. No:2003-1). New York: College Entrance Examination Board.

Engelhard, G. (1996). Evaluating Rater Accuracy in Performance Assessment. *Journal of Educational Measurement*. 33, 1, 56-70.

Field, A. (2005, 10 Aralık). Factor Analysis Using SPSS. 2 Şubat 2008 tarihinde http://www.sussex.ac.uk/Users/andyf/factor.pdf adresinden erişildi.

Fine, A. E. (1995). *An Investigation in The Use of Performance Assessment Tasks in Undergraduate Mathematics*. Yayınlanmamış doktora tezi. University of Oklohama: Oklahama.

Fleiss, J. L. (1971). Measuring Nominal Scale Agreement Among Many Raters. *Psychological Bulletin*. 76, 5, 378-382.

Frederikson, J. R. ve Collins, A. (1989). A Systems Approach to Educational Testing. *Educational Researcher*, 18, 27-32.

Gay, L. R. (1987). *Educational Research: Competencies for Analysis and Applications* (3. Basım). New York: MacMillan.

Goodwin, L. D. (2001). Interrater Agreement and Reliability. *Measurement in Psychical Education and Exercises Science,* 5 (1), 13-14.

Hambleton, R. K., Swaminathan, H. ve Rogers, H.J. (1991). Fundamentals of Item Response Theory. Sage Publications. The International Professional Publishers.

Hetherman, S. C. (2004). *An Application of Multi Faceted Rasch Measurement to Monitor Effectiveness of The Written Composition in English in The New York City Department of Education.* Yayınlanmamış doktora tezi. Teacher College, Colombia University, Colombia.

Howell, D. C. (2002). *Statistical Methods for Psychology*.(Fifth Edition). Thomson Learning Academic Research Center, USA.

Keats, J. A. (1997). Measurement in Educational Research. *Educational Research Methodology and Measurement an International Handbook.* (Second Editon). Editor Keeves, J. P. Cambridge, UK.

Kieffer, K. M. (1998). Why Generalizability Theory is Essential and Classical Test Theory is Often Inadequate? Paper presented at the annual meeting of the Southwestern Psychological Association. New Orleans, LA. USA.

Klein, S. P., Stecher, B. M., Shavelson, R. J., McCaffrey, D., Ormseth, T., Bell, R. M. ve diğerleri. (1998). Analytic versus Holistic Scoring of Science Performance Tasks. *Applied Measurement in Education*, 11, 121-137.

Lee, G. ve Frisbie, D. A. (1999). Estimating Reliability Under a Generalizability Theory Model for Test Scores Composed of Testlets. *Applied Measurement in Education*. 12, 3, 237-255.

Lee, Y. W. ve Kantor, R. (2003). Investigating Differential Rater Functioning for Academic Writing Samples: An MFRM Approach. Paper to be presented at the annual meeting of National Council on Measurement in Education held in Chicago, IL.

Lei, P., Smith, M. ve Suen, H. K. (2007). The Use of Generalizability Theory to Estimate Data Reliability in Single Subject Observational Research. *Psychology in Schools*. 44, 5, 433-439.

Linacre, J. M. (2007). *A User's Guide to FACETS. Rasch Model Computer Programs*. Chicago, IL.

Linacre, M. C. (1994). *Many-facet Rasch Measurement*. Chicago:MESA Press.

Linacre, J. M. (1989). *Many Facet Rasch Measurement*. Yayınlanmamış doktora tezi. University of Chicago, Chicago.

Linn, R. L., ve Gronlund, N. E. (2000). *Measurement and Assessment in Teaching*. Prentice Hall, Columbus, Ohio, USA.

Linn, R. L. (1980). Issues of Validity for Criterion- Referenced Measures. *Applied Psycological Measurement,* 4(4), 547-561.

MacMillian, P. D. (2000). Classical, Generalizability and Multifaceted Rasch Detection Interrater Variability in Large, Sparse Data Set. *Journal of Experimental Education*. 68, 2, 167-190.

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.

Mertler, C. A. (2001). *Designing Scoring Rubrics for Your Classroom*. Practical Assessment Research and Evaluation, 7 (25).

Meyer, G. J. (1999). Simple Procedures to Estimate Chance Agreement and Kappa for The Interrater Reliability of Response Segments Using The Rasch Comprehensive System. *Journal of Personality Assessment*, 72, 230-255.

Moon, S. Y. (1995). *Performance Assessment: Measurement Issues of Generalizability, Dependability of Scoring and Relative Information on Student Performance*. Yayınlanmamış doktora tezi. The Florida State University, Talhasse.

Moskal, B. M. (2000). Scoring Rubics: What, When and How?. *Practical Assessment, Research and Evaluation*, 7 (3).

Mushquash, C. and O'Connor, B. P. (2006). SPSS and SAS Programs for Generalizability Theory Analysis. *Behavior Research Methods*. 38 (3), 542-547.

Nakamura, Y. (2002). Teacher Assessment and Peer Assessment in Practice. *Educational Studies*, 44.

Nakamura, Y. (2000). Many Facet Rasch Based Analysis of Communicative Language Testing Results. *Journal of Communication Students*, V12, 3-13.

Nitko, A. J. (2001). *Educational Assessment of Students* (3. Baskı). Upper Saddle River, NJ: Merill.

PISA 2003 Projesi Ulusal Nihai Raporu (2005). Milli Eğitim Bakanlığı Eğitimi Araştırma ve Geliştirme Dairesi Başkanlığı.

Rogers, H. J. (1997). Multiple Choice Tests, Guessing. *Educational Research Methodology and Measurement an International Handbook.* (Second Editon). Editor Keeves, J. P. Cambridge, UK.

Shavelson, R. J. ve Webb, N. M. (1991). *Generalizability Theory*: A Primer. Sage Publications, USA.

Shavelson, R. J., Webb, N. M. ve Rowley, G. L. (1989). Generalizability Theory. *American Psychologist*. 44, 6, 922-932.

Smith, V. E. and Kulikowich, M. J. (2004). An Application of Generalizability Theory and Many Facet Rasch Measurement Using A Complex Problem-Solving Skills Assessment. *Educational and Psychological Measurement*, 64, 617-639.

Sönmez, V. (2003). *Program Geliştirmede Öğretmen El Kitabı*. Ankara: Anı Yayıncılık.

Sub, A., Valiga, M. J. ve Goa, X. (1997). Using Generalizability Theory to Assess the Reliability of Student Ratings of Academic Advising. *Journal of Experimental Education*. 65, 4, 367-379.

Sudweeks, R. R., Reeve, S. and Bradshaw, W. S. (2005). A Comparison of Generalizability Theory and Many Facet Measurement in An Anlaysis of College Sophomore Writing. *Assessing Writing*. 9, 239-261.

Teresa, S. A. (1997). The Generalizability of Scoring TIMSS Open Ended Items. *Paper presented at the Annual Meeting of the American Educational Research Association*. New Orleans, LA.

TIMSS, (2008). TIMSS & PIRLS International Study Center. 20 Şubat 2008 tarihinde http://timss.bc.edu/ adresinden erişildi.

TIMSS 199 Türkiye Raporu (2003). Üçüncü Uluslar arası Matematik ve Fen Bilgisi Çalışması. Milli Eğitim Bakanlığı Eğitimi Araştırma ve Geliştirme Dairesi Başkanlığı.

Turner, J. (2003). *Examining on Art Portfolio Assessment Using A Many Facet Rasch Measurement Model*. Yayınlanmamış doktora tezi. Boston College, Boston.

Wiggins, G. (1998). *Educative Assessment: Designing Assessment to Inform and Improve Student Performance*. San Francisco: Jassey-Bass Publishers.

Wolf, R. M. (1997). Rating Scale. *Educational Research Methodology and Measurement an International Handbook.* (Second Editon). Editor Keeves, J. P. Cambridge, UK.

Wright, B. and Stone, M. (1979). *Best Test Design: Rasch Measurement*. Chicago:MESA Press.

Yelboğa, A. (2007). *Klasik Test Kuramı ve Genellenebilirlik Kuramına Göre Güvenirliğin Bir İş Performansı Ölçeği Üzerinde İncelenmesi.* Yayınlanmamış doktora tezi. Ankara Üniversitesi, Ankara.

EK 1

MATEMATİK PERFORMANSININ ÖLÇÜLMESİNE İLİŞKİN ARAŞTIRMA SORULARI

1) Aynı koşu pistinde, Aliye ile Kadriye aynı anda koşmaktadırlar. Kadriye pisti üç tur koştuğunda Aliye dört tur koşabilmektedir. Kadriye pisti 12 tur koştuğunda Aliye pisti kaç tur koşar?

Çözüm:

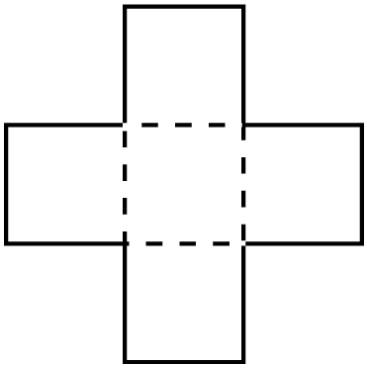
2) Can'ın üç testten aldığı puanlar 78, 76 ve 74 tür. Can'ın ortalama puanı kaçtır?

Çözüm:

3) Bir arabanın yakıt deposu 45 litre yakıt almaktadır. Araba her 100 km yol aldığında, 8,5 litre yakıt tüketmektedir. Bir depo dolu yakıtla 350 km'lik bir yolculuğa çıkılmıştır. Yolculuğun sonunda depoda ne kadar yakıt kalmıştır ?

Çözüm:

4) Şekil eşit alanlı 5 kare içermektedir. Tüm şeklin alanı 245 cm^2 dir.



a) Bir karenin alanını bulunuz.

Çözüm:

b) Bir karenin bir kenarının uzunluğunu bulunuz.

Çözüm:

c) Tüm şeklin çevresi kaç cm dir?

Çözüm:

5)

9	1	4	5

Yukarıdaki dört rakam en büyükten en küçüğe doğru dört rakamlı bir sayı oluşturmak için düzenlenecektir. Sonra aynı dört sayı en küçükten en büyüğe doğru dört rakamlı diğer bir sayı oluşturmak için düzenlenecektir. Elde edilen iki sayı arasındaki fark kaçtır?

Çözüm:

6) 20 cm uzunluğundaki ince bir kablo dikdörtgen şekline getiriliyor. Eğer bu dikdörtgenin eni 4 cm ise boyu kaç cm dir?
Çözüm:

7) Alican bir koşu yarışını 49,86 saniyede bitirdi. Betül aynı yarışı 52,30 saniyede bitirdi. Yarışı bitirmek Betül'ün ne kadar daha fazla zamanını almıştır?

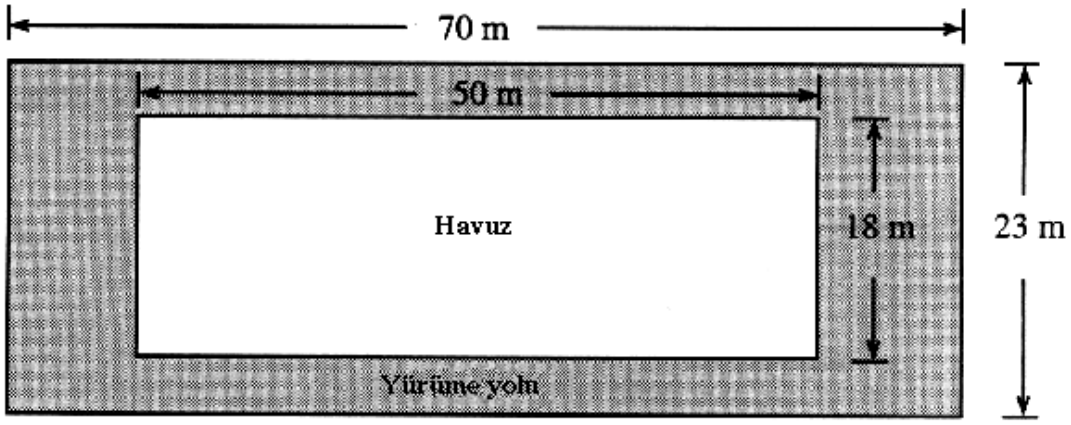
Çözüm:

8) Bir öğretmenin ve bir doktorun 45 er kitabı vardır. Eğer öğretmenin

$\frac{4}{5}$ i kitaplarının ve doktorun $\frac{2}{3}$ ü kitaplarının roman ise, öğretmenin romanları doktorun romanlarından ne kadar daha fazladır?

Çözüm:

9) Dikdörtgen biçimli bir yüzme havuzunun etrafında şekilde görüldüğü gibi bir yürüme yolu vardır.



Havuzun etrafındaki yürüme yolunun alanı nedir ?

Çözüm:

10) Bir sınıfta 86 öğrenci vardır ve kızların sayısı erkekelerden 14 fazladır. Sınıfın kaç erkek ve kaç kız öğrencisi vardır?

Çözüm:

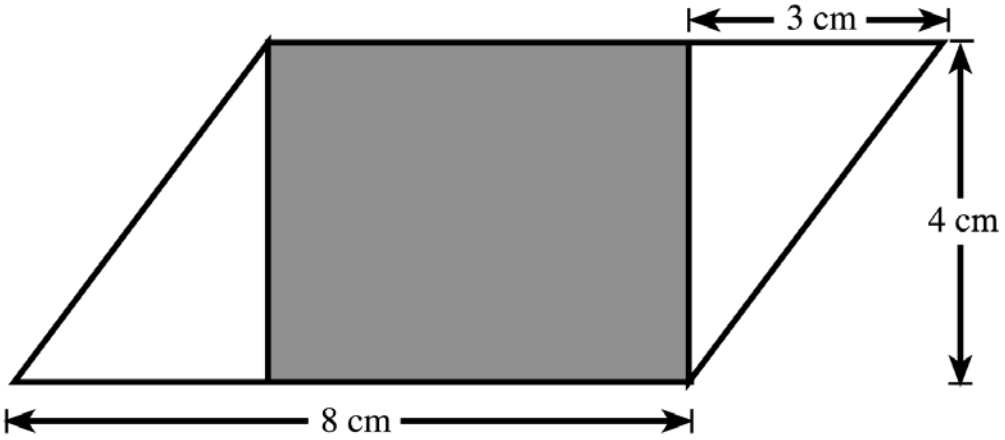
11) Bir kırtasiyeci 140 tane kalemi iki farklı türdeki kutulara koyarak paketlemektedir. Kutulardan biri 8, diğeri 12 kalem almaktadır. Kutuların hepsi de dolu ve her iki kutudan da eşit sayıda bulunmaktadır.

a) Bu kırtasiyecinin 12 kalem alan kaç kutusu vardır?

Çözüm:

b) 8 kalem alan kutuların tümünde bulunan kalem sayısı kaçtır?
Çözüm:

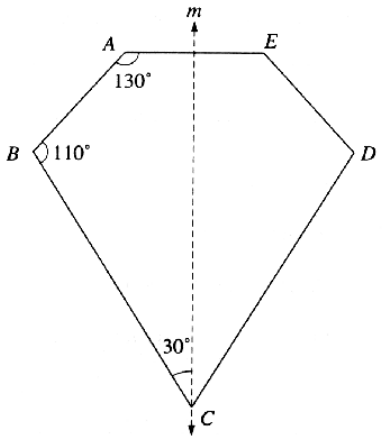
12) Aşağıdaki şekil, bir paralel kenar içindeki taranmış bir dikdörtgeni göstermektedir.



Taralı dikdörtgenin alanı kaç cm^2 dir?

Çözüm:

13) m doğrusu ABCDE şekli için bir simetri eksenidir.
BCD açısının ölçüsü kaçtır?



Çözüm:

14) 7 nin 13 e oranı x in 52 ye oranı ile aynı ise x kaçtır?

Çözüm:

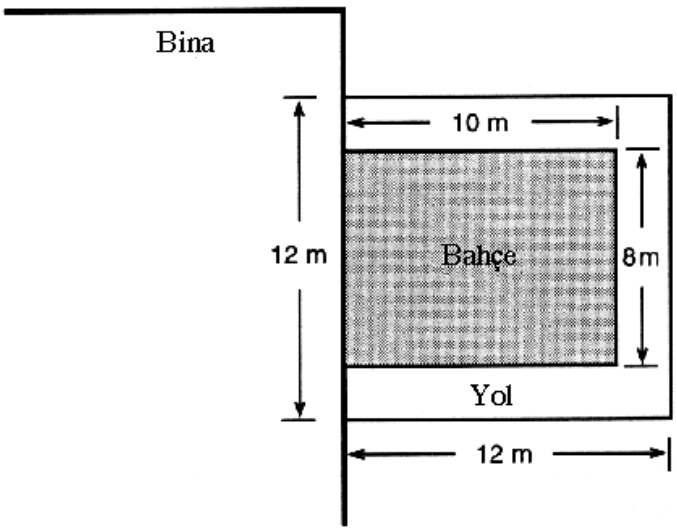
15) 3 000 metreyi 8 dakikada koşan bir koşucunun ortalama hızı kaç metre/dakika dır?

Çözüm:

16) Bir sayının 4 katı 48 ise dörtte biri kaçtır?

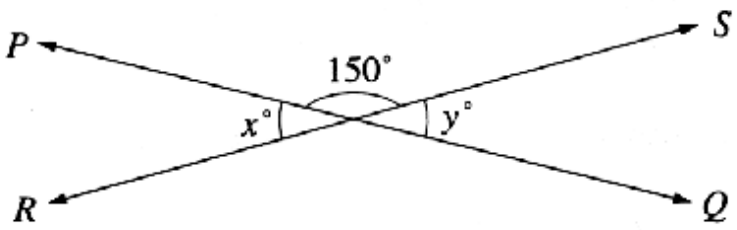
Çözüm:

17) Bir binanın yanındaki dikdörtgen şeklindeki bir bahçenin üç kenarında aşağıda gösterildiği gibi bir yol bulunmaktadır. Yolun alanı (**taralı olmayan alan**) ne kadardır?



Çözüm:

18) Şekilde , PQ ve RS kesişen doğrulardır. Buna göre, $x + y$ kaçtır?

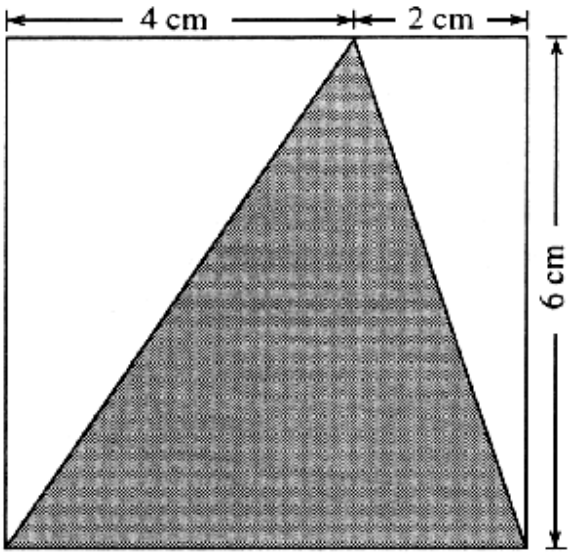


Çözüm:

19) Bir sınıftaki öğrencilerin 16 sı gözlüklü; 14 ü gözlüksüzdür. Sınıfta gözlüklü olan öğrenciler sınıftaki tüm öğrencilerin kaçta kaçıdır ?

Çözüm:

20) Aşağıdaki şekil, bir kare içindeki taralı bir üçgeni göstermektedir. Taralı üçgenin alanı kaç cm^2 dir?



Çözüm:

21) 3 000 tane ampul arasından 100 tane ampul gelişigüzel seçiliyor ve test ediliyor. Seçilen ampullerden 5 tanesi kusurluysa, tamamında kaç ampulün kusurlu olması beklenir?

Çözüm:

22) Canan 24 sayılık bir dergi sipariş etmeyi planlıyor. İki dergi ile ilgili aşağıdaki ilanları okuyor.

Gençlik Dergisi	**Genç Haber**
24 sayı İlk dört sayı ÜCRETSİZ kalanların her biri 3 YTL	**24 sayı** İlk altı sayı ÜCRETSİZ kalanların her biri 3.5 YTL

24 sayılık en ucuz dergi hangisidir? Ne kadar daha ucuzdur?
Çözüm:

23) Bir torbada, 20 tane beyaz, 30 tane siyah ve 60 tane de kırmızı bilye bulunmaktadır. Kırmızı bilyelerin, torbadaki toplam bilye oranı nedir?

Çözüm:

24) Bir dergiden, her hafta 7 000 adet satıldığına göre, yılda kaç adet dergi satılmaktadır?

EK 2

Matematik Performansının Ölçülmesine ilişkin Bütünsel Dereceli Puanlama Anahtarı

5 Puan (Örnek cevap)	Tam ve tutarlı cevaplandırılmış. Problemin matematiksel fikir ve süreçlerini anladığını göstermiştir.
4 Puan (Tatmin edici cevap)	Cevap tam, ancak yeterli açıklama yok.
3 Puan (Başlar ama problemi sonlandıramaz)	-Probleme uygun başlar ancak sonu yoktur. -Problemin matematiksel fikir ve süreçlerini anladığını göstermiş, ancak basit matematiksel hatlar yapmış olabilir.
2 Puan (Uygun olmayan cevap)	-Problemin matematiksel fikir ve süreçlerini anladığını göstermemiş olabilir. -Cevap, problemi çözmek için uygun olmayan bir stratejiyi yansıtıyor olabilir. -Temel hesaplama hataları yapmış olabilir.
1 Puan	-Sadece doğru cevap var. Hiçbir açıklama, problemi anladığına dair bir ifade bulunmamaktadır.
0 Puan	-Boş bırakılmış. -Cevapla ilgisi olmayan, anlamsız, karalama şeklinde ifadeler var.

Emeğiniz, ilginiz ve desteğiniz için çok teşekkür ederim.
Neşe Güler.

EK 3

VERİ GİRİŞİ TABLO ÖRNEĞİ

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Y1																								
Y2																								
Y3																								
Y4																								
Y5																								
Y6																								
Y7																								
Y8																								
Y9																								
Y10																								
Y11																								
Y12																								
Y13																								
Y14																								
Y15																								
Y16																								
Y17																								
Y18																								
Y19																								
Y20																								
Y21																								
Y22																								
Y23																								
Y24																								
Y25																								
Y26																								
Y27																								
Y28																								
Y29																								

EK 4

24 MADDENİN FAKTÖR ANALİZİNE GÖRE FAKTÖR YÜKLERİ

Madde No:	Faktör-1	Faktör-2	Faktör-3	Faktör-4	Faktör-5
1					0.466
2		0.636			
3	0.585				
4	0.658				
5	0.484	0.429			
6	0.529				
7	0.452				
8	0.578				
9	0.650				
10	0.509				-0.451
11	0.668				
12	0.662				
13	0.602				
14	0.578				
15	0.526				
16	0.496			0.545	
17	0.683				
18	0.618				
19	0.692				
20	0.761				
21	0.643				
22	0.589		-0.565		
23	0.663		-0.451		
24	0.560				