

Problem Durumu: Ölçme, bir özelliğin gözlenerek gözlem sonuçlarının sayı ya da sembollerle ifade edilmesi olarak tanımlanırken, değerlendirme daha geniş kapsamlı bir süreç olup; ölçme sonuçları, 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 güvenilirliğinin ve geçerliğinin olabildiğince yüksek olması istenir. Güvenirlik, ölçme sonuçlarının tesadüfi hatalarından arınık olma derecesi olarak tanımlanmaktadır. Güvenirlik, ölçümlerin zaman içerisindeki tutarlılıklarının derecesidir. Güvenirliğin derecesi genellikle bir katsayı ile ifade edilir. Bu katsayı "0" (güvenilir değil) ile "1" (mükemmel güvenirlik) arasında değişen değerler alır ve ölçme sonuçlarının tesadüfi hatalardan ne derece arınık olduğunu gösterir. Öğrenci başarısını açık-uçlu sorularla ölçülmesinde, güvenirlik en zayıf halka olarak düşünülmektedir. Puanlayıcı, açık-uçlu sorulara verilen cevapların puanlanmasında güvenirligi düşüren önemli bir hata kaynağı olmakla birlikte, görev ya da madde gibi diğer faktörler de en az o kadar önemli diğer hata kaynaklarıdır. Bu nedenle sadece puanlayıcılar arası tutarlılığı sağlamak puanlamanın güvenirligi için yeterli olmamakta, diğer hata kaynaklarının da güvenirligin hesaplanmasında göz önünde bulundurulması gerekmektedir. Ancak güvenirligin hesaplanması için kullanılan tüm yöntemler, tüm hata kaynaklarıyla aynı anda ele almamaktadır. Açık-uçlu sorularla yapılan ölçmelerin güvenirligi, ölçmenin üç temel kuramı olan klasik test kuramı, madde tepki kuramı ve genellenebilirlik kuramına dayalı yöntemlerle çalışılmaktadır.

Araştırmanın Amacı: Bu araştırmayla, güvenirligin belirlenmesinde klasik test kuramı klasik test kuramına dayalı ölçme yöntemi ve çok değişkenlik kaynaklı Rasch modeli (ÇDKRM) yaklaşımları kullanılarak birbirleriyle ve kendi içlerinde tutarlılıkları sınanarak benzer ölçme durumları için kuramsal bir katkı sağlanabileceği düşünülmektedir. Matematik eğitimcileri, matematikle ilgili başarının ölçülmesinde subjektif ölçme araçları kullandıklarında, öğrencilerinin matematiksel bilgilerine ilişkin verdikleri puanların ne kadar güvenilir olduğunu bilmek isterler. Bununla birlikte, matematik başarısının ölçülmesinde en etkili olan değişkenlik kaynağının ne olduğunu ve ölçme hatasını en aza indirmek için ölçmenin nasıl yapılması gerektiğini bilmeye ihtiyaç duyarlar. Aynı zamanda bu noktalara temas edilerek matematik eğitimcilerini bu konularda aydınlatmak, çalışmanın bir diğer amacıdır.

Araştırmanın Bulguları: Çalışmada kullanılan 18 matematik sorusunun Cronbach α güvenirlilik katsayıları birinci ve ikinci puanlayıcılar için 0.91 ve üçüncü ve dördüncü puanlayıcılar için 0.92 olarak bulunmuştur. Elde edilen bu katsayılara göre, maddelerin matematik başarısını birbirleriyle

tutarlı bir şekilde ölçtükleri yorumu yapılabilir. Yapılan uygulamada, dört farklı puanlayıcının aynı koşullar altında her bir öğrenciyi puanlamasına ilişkin elde edilen puanlar arasındaki tutarlılık derecesi, Kendall'ın uyum katsayısı ile analiz edilmiş ve sonuç olarak 18 madde için 0.52 olarak bulunmuştur ($\chi^2=315.16$, $sd=3$, $p=0.00$). Ayrıca, her bir puanlayıcının verdiği puanlar ile diğer bir puanlayıcının verdiği puanlar arasındaki korelasyon katsayıları hesaplanmıştır. Puanlayıcıların 18 madde üzerinden verdikleri puanlar arasındaki korelasyonlar 0.90 ile 0.97 arasında değişen oldukça yüksek değerlere sahiptir. Elde edilen bu değerler puanlayıcılar arası uyum olduğu yorumunu destekler niteliktedir. Çalışmada yer alan dört puanlayıcının puanları ortalamaları arasındaki farklılık, ilişkili örneklem üzerinde tek değişkenli varyans analizi ile test edilmiş ve istatistiksel olarak anlamlı bir fark olduğu sonucuna varılmıştır ($F=13.801$, $p=0.00$). Bu sonuç üzerine yapılan post-hoc çalışmasıyla puanlayıcıların puanları ortalamalarının ikili karşılaştırılması sonucunda birinci puanlayıcı hariç diğer tüm puanlayıcıların verdikleri puanların ortalamaları arasında manidar bir farklılık görülmemiştir. ÇDKRM için FACET analizi sonuçlarında, 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 puanlarından 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, FACET analizi için model veri uyumunu sağladığı yorumu yapılabilir.

Ölçme sonucu, 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 güvenilirliği 0.95 olarak yeterince yüksek bulunmuştur. Hem iç uyum (infit) istatistikleri hem de dış uyum (outfit) istatistikleri verilerin modele çok iyi uyum gösterdiğine işaret etmektedir. Öğrencileri puanlayan dört puanlayıcının puanlamasındaki katılık/cömertliklerine ilişkin logit değerlerin -0.10 ile 0.22 arasında değiştiği görülmektedir. Ayırma indeksine bakıldığında 10.59 gibi oldukça yüksek bir değerle karşılaşılmaktadır. Bu değer puanlayıcılar arası farklılık olduğunun bir göstergesidir. Ayırma indeksi güvenilirliği de 0.99 olarak elde edilmiştir. Hem iç uyum hem de dış uyum değerleri, verinin genel anlamda modele çok iyi uyum gösterdiğine işaret etmektedir. 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. Maddelerin güçlük düzeyleri -0.46 ile 0.48 arasında değişmektedir. Maddelere ilişkin ayırma indeksi 8.36 ve güvenilirlik 0.99 olarak elde edilmiştir. Hem iç uyum hem de dış uyum değerleri, verinin genel anlamda modele çok iyi uyum gösterdiğine işaret etmektedir. Bu değerler, verilerin ölçüm için uygun olduğunu işaret etmektedir.

Sonuç ve Öneriler: 2007 yılında uygulanan matematik başarısının ölçülmesinde yer alan maddelerin matematik başarısını oldukça tutarlı bir şekilde ölçtüğü sonucuna varılmıştır. Çalışmada yer alan dört puanlayıcının puanlamadaki katılık/cömertlik düzeylerinin birbirinden farklı olduğu görülmekle birlikte puanlayıcıların puanları sıralaması paralellik göstermekte olup, birbirleriyle tutarlı olacak şekilde puanlama yaptıkları sonucuna varılmıştır.

ÇDKRM'ne göre, FACET analizi ile modelin veri için uygun olduğu sonucuna varılmıştır. FACET analizi ile ayırma indeksi hesaplanmış ve bu değerin oldukça yüksek olduğu belirlenmiştir. Öğrenci değişkenlik boyutu için ayırma indeksi güvenilirliği; KR-20, Cronbach alfa ve genellenebilirlik katsayısına benzer yorumlanmaktadır. Buna göre, matematik puanlarının ÇDKRM'ne göre iç tutarlılık anlamında güvenilirliğinin yüksek olduğu sonucuna varılmıştır. Dört puanlayıcının, 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. FACETS analiziyle maddelerin güçlük dereceleri açısından birbirlerinden farklı oldukları sonucuna varılmıştır.

ÇDKRM, tek bir çalışma ile her bir değişkenlik kaynağına ilişkin farklı güvenilirliklerin 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 ÇDKRM'nin kullanılması uygun olabilir. Özellikle birden fazla puanlamanın yapıldığı durumlarda güvenilirliğin kestirilmesinde, klasik test kuramının yanı sıra ÇDKRM'inin kullanılması önerilebilir. Farklı ölçme durumlarında iki kuramın birlikte kullanıldığı benzer çalışmaların yapılmasının güvenilirlik çalışmalarına katkı sağlayacağı düşünülmektedir.