PSİKOLOJİDE İYİ BİR TESTİN ÖZELLİKLERİ*

*Çev: Nalan SANLI***

Ancak belirli özelliklere sahip bir test, iyi bir psikolojik test olarak tanımlanabilir. Test, enazından bir eşit aralıklı ölçek olmalı, daha sonra geçerli, güvenilir, ayırtedici gücü yüksek, iyi hazırlanmış normlara sahip bulunmalı ya da Rasch veya benzer modellere yüksek bir kesinlikle uymalı veya testi alacak kişilere uygun olarak biçimlendirilmiş olmalıdır.

Burada, anlamlı ve yaratıcı bir test geliştirme ile bu özelliklerin nasıl oluşturulabileceğini göstermek amaçlanmaktadır. Bununla beraber, bu yapılmadan önce bütün bu kavram ve terimlerin tartışılması, tanımlanması ve tam olarak anlaşılması, testlerin yalnızca uygun bir biçimde uygulanması için değil, aynı zamanda uygun bir biçimde kullanılması için de gereklidir.

Psikolojik testlerin bu özelliklere sahip olmasını gerekli kılan en önemli neden, ölçmenin doğruluğunu ve kesinliğini arttırma çabalarıdır. Bu özellikler, böyle bir ölçmenin bilimin mutlaka aranılması gereken koşulları oldukları için de istenilmektedir.

Ölçek Türleri

Ölçeklerin çok farklı düzeyleri vardır ve bunlar hiyerarşik olarak sıralanmıştır. Bunlar, aşağıdaki gibi basitten başlayarak karmaşığa doğru sıralanır:

Sınıflama: Bu ölçek nesneleri sadece sınıflandırır. Kadın-Erkek sınıflama ölçeğine göre yapılmış bir sınıflamadır.

Sıralama: Bu ölçekte nesneler boy ve ağırlıkta olduğu gibi sıralanır. Oldukça kaba bir ölçektir. Çünkü sıralar arasındaki farklar gözardı edilmiştir.

---

* Kline, Paul (1986). A Handbook of Test Construction: Introduction to Psychometric design. "The Characteristics of Good tests in Psychology." Sayfa 1-23. Methoen 8 Co. Ltd.

** Araş. Gör. A.Ü. Eğitim Bilimleri Fakültesi. Eğitimde Psikolojik Hizmetler Bölümü Ölçme Değerlendirme Anabilim Dalı.

Eşit aralıklı: Burada ölçek noktaları arasındaki farklar, ölçeğin bütün noktalarında eşittir. Eşit aralıklı ölçekler lineer olarak dönüştürülebilir. Böylece puanların diğer ölçeklere dönüştürülmesine ve puanların karşılaştırılmasına olanak sağlamış olur. Ayrıca birçok istatistiksel yöntem ölçmenin eşit aralıklı ölçek ile yapıldığını kabul eder.

Eşit Oranlı Ölçek: Eşit aralıklı ölçekler belirlenen diğer özelliklere ek olarak anlamlı bir "0" sıfır noktasına sahiptirler. Eşit oranlı ölçek kullanımına olanak sağlayan test geliştirme yöntemleri olmasına rağmen, bu ölçek türü birçok psikolojik değişken için belirgin bir problem oluşturur.

Bu dört farklı türdeki ölçeklerin incelenmesi psikolojik testleri geliştiren kişilerin ideal olarak eşit oranlı ölçek üretmeyi amaç edinmeleri gerektiğini ortaya koymaktadır. Bunun başarılamadığı durumlarda sonuçların herhangi bir istatistiksel analize uygulanabilmesine olanak sağlayacak eşit aralıklı ölçek olması istenir. Testlerin geçerlilik çalışmaları kaçınılmaz olarak bu tür analizleri kapsadığı için ve bu analizler, puanların nicelendirilmesi ile yapıldığı için, ki bunlar: psikolojik testlerin diğer değerlendirme biçimlerine olan üstünlüklerini sağlar, ölçeğin en az eşit aralıklı olması gerektiği gerçeği oldukça açıktır. Aslında Brown'un (1978) belirttiği gibi birçok psikolojik test, eşit aralıklı ölçeklemeye yaklaşır ve test puanlarını eşit aralıklı ölçek kullanılmış gibi ele alır.

Güvenirlilik

Psikometride "güvenirlilik" kavramı iki farklı anlam taşır. Bir test kendi içinde tutarlı ise testin güvenilir olduğu söylenebilir. Test aynı gruba tekrar uygulandığında her öğrenci için aynı puanı veriyorsa (öğrencilerin değişmediği düşünülerek) bu durumda da testin güvenilir olduğu söylenebilir. Zamana dayalı olan bu güvenirlilik, test-tekrar test güvenirliliği olarak bilinir.

İç Tutarlılık Güvenirliliğinin Önemi ve Anlamı

Psikometristler iç tutarlılığı yüksek testler geliştirmek isterler ve bunun nedeni de oldukça açıktır. Eğer testin bir bölümü bir değişkeni ölçüyor ve testin diğer bölümleri bu bölümle tutarlı değilse, test bu değişkeni ölçemez. Öyle ise bir testin geçerli olabilmesi için (ölçmek istediği özelliği ölçebilmesi) tutarlı-sürekli olması gerekir. Bu nedenle psikometristler iç tutarlılık katsayısı üzerinde önemle durmaktadırlar. Gerçekte, genel psikometrik yaklaşım yüksek güvenirliliğin geçerliliğin ön koşulu olduğu yönündedir. Bu konudaki tek aykırı görüş Cattell'e aittir. Cattell'e göre yüksek bir iç tutarlılık, herhangi bir test maddesi ölçmeye çalıştığımız özellikten daha az ya da daha dar kapsam içermesi gerektiğinden gerçekte geçerliliğe tezattır. Böylece, eğer bütün maddeler oldukça tutarlı ise aralarındaki korelasyonda yüksektir. Bu durumda güvenilir bir test sadece çok küçük bir varyansın dar bir değişkenini ölçebilmektedir.

1. Cronboch Alpha'sının madde korelasyonları ile birlikte artması ve 2. herhangi bir çok değişkenli yordama çalışmasında testler ve ölçütler arasındaki maksimum çoklu korelasyonun (test puanları ve toplam puan durumu da olduğu gibi), değişkenler arasında korelasyon olmadığı zaman elde edilmesi, bu görüşü destekler niteliktedir.

İki değişkenin mükemmel bir korelasyon gösterdiği durumlarda, bunlardan birisinin yeni bir bilgi veremeyeceği ortadadır. Bu durumda Cattall'e göre maksimum geçerlilik, test maddelerinin birbiri ile tam korelasyon göstermemesi fakat ölçüt ile pozitif bir korelasyon göstermeleri ile mümkündür. Böyle bir test, düşük iç tutarlılık güvenirliliğine sahip olabilir. Cattell teorik olarak haklı olabilir. Fakat hiçbir test geliştiricisi madde yazımı sırasında, maddelerin birbirleri ile ilişki göstermemesi fakat ölçüt ile ilişki göstermesi yönünde manüplasyonda bulunmaz. Barret ve Kline (1982) bu yönde bir girişimde bulunulmuş olan, Cattell'in geliştirdiği kişilik testini (16 PF) incelemişler fakat bu yöntemi pek başarılı bulmamışlardır. Bütün bu görüşlere rağmen, genel psikometrik bakış açısına göre pratikte geçerli testler aynı zamanda tutarlılığı yüksek olan testlerdir.

Test-Tekrar Test Güvenirliliği

Eğer bir test, farklı uygulama oturumlarında bir kişi için aynı puanı vermekte başarısız oluyorsa (verilen özelliğin değişmediği durumlarda) test, iyi bir test olamaz. Bu nedenle test-tekrar test güvenirliliği çok önemlidir ve belirlenmesi de oldukça basittir. Bir gruba, iki farklı zamanda yapılan uygulamalardan elde edilen sonuçlar ilişkilendirilir. Test güvenirliliği için istenen minimum katsayı 0.70 dir. Guilford'un (1956) da belirttiği gibi bunun altındaki durumlarda, testin kişilerle kullanılması doyurucu sonuçlar vermez. Çünkü elde edilen puanların standart hatası öylesine büyük olur ki, puanları yorumlanması yanıltıcı sonuçlar verir. Test-tekrar test güvenirliliğini hesaplamak kolay olmakla birlikte iki uygulama arasında geçen sürenin çok kısa olması nedeni ile katsayının yapay olarak artması ihtimaline ve seçilen örneklemin, testin uygulanacağı populasyonu temsil edici olmasına dikkat edilmelidir.

Paralel Form Güvenirliliği

Eşdeğer (paralel) form güvenirliliğini sağlamak için maddelerin eşdeğer yada paralelleri oluşturulur ve grup, izleyen uygulamalarda tamamen farklı bir test almış olur. İki formun gerçekten eşdeğer olduğunu göstermede bazı problemler olmasına rağmen, bir testin paralel formunun bulunması yararlıdır.

Geçerlilik

İyi bir psikolojik testin bir diğer önemli özelliği geçerliliktir. Bir test, ölçmek istediği özelliği ölçebiliyorsa geçerlidir. Fakat bu tanım, geçerliliğin anlamını tam olarak ortaya koymayıp aksine, ölçmek istediğimiz özelliği ölçüp ölçmediğimizi nasıl bileceğimiz ile ilgili yeni soruların oluşmasını sağlamaktadır. Aslında bir testin geçerliliğini gösterecek birçok farklı yol vardır. Bunlardan herbiri geçerliliğin farklı bir boyutuna katkıda bulunur.

Görünüş (Yüzeysel) Geçerliliği

Eğer bir test ölçmek istediği özelliği ölçüyor görünüyorsa (özellikle testi alan kişilere) testin görünüş geçerliliğine sahip olduğu söylenebilir. Görünüş geçerliliği, gerçek geçerlilik ile ilgili bir anlam taşımaz, sadece yetişkinler, görünüş geçerliliğine sahip olmayan testlere genellikle katılımda bulunmadıkları ya da böyle bir testi saçma buldukları için önemlidir. Öyleyse, bu geçerlilik çeşiti, testi alan kişilerin katılımını, yardımını sağlamak için gereklidir.

Halihazır Geçerlilik

Bu geçerlilik, testin diğer testlerle olan korelasyonu yoluyla değerlendirilir. Bir zeka testinin halihazır geçerliliğini sağlamaya çalışıyorsak, bu testin, daha önceden geçerli olduğu bilinen bir başka testle korelasyonunu bulabiliriz. Bu örnek, halihazır geçerliliğin içinde bulunduğu çelişkiyi açıkça göstermektedir. Zaten ölçüt olabilecek derecede yeterli ve geçerli bir test varsa, yeni bir testin geçerliliğinin sağlanması belkide yararsız bir iş olarak görülebilir. Diğer testte bulunmayan bazı olumlu özelliklere sahip olmadığı sürece, gerçekten de öyle olacaktır. Çok kısa, uygulamanın kolay, puanlamanın çabuk olması gibi özellikler, diğer ölçüt olarak kullanılan testlerin varlığına ragmen, yeni bir testin geliştirilmesini haklı hale getirebilir. Diğer taraftan, iyi bir ölçüt testin olmadığı ve yeni testin ilk olarak bir girişimi başlattığı durumlarda, halihazır geçerlilik çalışmaları çok güç olmaktadır.

Genel olarak halihazır geçerlilik aynı değişken üzerinde daha zayıf bir testin var olduğu ve bu yeni testle, bu konuda bir gelişme sağlandığı durumlarda yararlıdır. Böyle durumlarda, halihazır geçerlilik çalışmalarından anlamlı fakat, çok yüksek olmayan korelasyon katsayıları beklenir. Daha açık olarak söylemek gerekirse, halihazır geçerlilik, geçerliliğin tamamen doyurucu bir boyutu değildir. Bir testi, geçerli bir test olarak kabul edebilmemiz için, buna ek olarak daha farklı ve güçlü kanıtlara ihtiyacımız vardır. Testin neyi ölçtüğü kadar neyi ölçmediğini ortaya koymakta önemlidir. Bir başka deyişle testin oldukça farklı değişkenleri ölçen başka testlerle ilişki göstermemesi gerekir.

Yordama Geçerliliği

Bir testin yordama geçerliliğini sağlamak için, uygulama sonrası elde edilen puanlarla, daha sonraki bir zamanda elde edilen ölçüt arasındaki ilişki bulunur. Örneğin bir zeka testinin yordama geçerliliği, 11 yaşında testten elde edilen puanla, o kişinin üniversitedeki notları ilişkilendirilerek gösterilebilir. Bir çok psikometrist yordama geçerliliğini, bir testin etkililiğini gösteren en güçlü kanıt olarak görmektedir.

Test geçerliliğine ilişkin bu yaklaşımdaki en büyük güçlük anlamlı bir ölçütün bulunamamasındadır. Zeka testlerinde olduğu gibi, verilen zeka kavramına bağlı olarak gelecekdeki akademik başarıyı hatta çalışılan işte kazanılan parayı ölçüt olarak kullanmak anlamlıdır. Fakat bu ölçüt ile ilişkili olabilecek zekadan başka değişkenler olduğu için, zeka testi ile ölçüt arasındaki korelasyonların sadece orta derecede olması beklenebilir. Ayrıca zeka, belkide yordama geçerliliği çalışmalarının anlamlı şekilde yapılabildiği en kolay değişkendir. Fakat birçok değişken için yordama geçerliliği çalışması yapmak güçtür. Örneğin Cattell'in C faktörü, benlik gücü (ego strenght), en yaratıcı araştırmalar için bile ciddi bir test etme sorunu oluşturur. Bunlara ek olarak örneğin, varyansın homojenliğinden kaynaklanan korelasyon katsayısının düşmesi gibi istatistiksel güçlükler de vardır.

Arttırıcı (İncremental) ve Ayırt Edici Geçerlilik (Discriminational)

Vernon (1950) tarafından tartışılan bu iki terimin kısaca açıklanması gerekir. Arttırıcı geçerlilik, bir test bataryasındaki testlerden birinin ölçüt ile düşük korelasyon göstermesi fakat bataryadaki diğer testlerle binişiklik göstermemesi durumlarında söz konusudur. Bu durumda, test ölçüt açısından seçim için arttırıcı geçerliliğe sahiptir. Ayırt edici geçerlilik belkide en iyi şekilde, ilgi testlerinde gösterilebilir. Bu testler üniversite başarısı ile sadece orta derecede ilişki gösterir fakat farklı kişiler için farklılaşır. Böylece, bu testlerin akademik performans için ayırtedici geçerliliğe sahip oldukları söylenebilir. Diğer taraftan IQ testleri üniversite notları ile daha büyük korelasyon gösterir fakat kişiler arasında ayırım yapamaz.

Özet olarak arttırıcı ve ayırt edici geçerlilikler, seçme amacı ile kullanılan testlerin etkinliğini ortaya koyan yararlı göstergelerdir.

Kapsam Geçerliliği

Bu terime daha çok başarı testlerinde başvurulur. Eğer bir testin maddelerinin, test edilen kişinin bütün yönlerini yansıttığı gösterilebilirse, verilen yönergeler oldukça açık olmak koşuluyla, test kendiliğinden geçerlidir. Bu, test maddelerinin görünüşü ile ilgili olan görünüş geçerli-

liği değildir. Eğer bir matematik testinde parantez içindeki terimleri çarpma yeteneğini sınamak istiyorsak ve testte (y+2k)(2y–3x)=? gibi maddeler varsa, bu maddelerin geçerli olmadığını söylemek oldukça zordur. Açıkça görülüyor ki, kapsam geçerliliği, matematikte olduğu gibi sadece ele alınan konunun açık, anlaşılır olduğu testler için kullanılır.

Yapı Geçerliliği

Bu kavram ilk olarak Crombach ve Meehl (1955) tarafından tanıtılmıştır. Bir testin yapı geçerliliğini gösterebilmek için, testin ölçmeyi amaçladığı değişkeni (yapıyı) olabildiğince açık olarak tanımlamak gerekir. Yapı geçerliliği, ölçülmek istenen değişken hakkında bütün bildiklerimizin ışığında testten elde edilecek puanlar ile ilgili hipotezler kurmak yoluyla oluşturulur. Böylece yapı geçerliliği daha önce tartışılan geçerlilik ile ilgili tüm yaklaşımları da kapsar. Bunu en iyi şekilde bir örnek kullanarak açıklayabiliriz.

Kline (1978) tarafından geliştirilen "Sözel Kötümserlik Envanteri"nin (The Oral Pessimism Questionnaire-OPQ) yapı geçerliliğinin oluşturulmasında test edilecek hipotezleri ele alarak inceleyelim.

1. OPQ, diğer sözel testlerle pozitif fakat orta derecede korelasyon göstermelidir. (Çok iyi testler olmadıkları için)

2. Sözel kötümserlik sendromu, tanımı gereğince nörotizm ile orta derecede bir korelasyon göstermelidir.

3. Cattell'in 16 PF faktörleri, Sözel Kötümserlik sendromuna benzer ölçekleri kapsamadığı için, OPQ ile aralarında korelasyon olmamalıdır.

4. OPQ bir kişilik testi olduğu için yetenek ya da motivasyon ile ilişkili değişkenler ile anlamlı bir korelasyon göstermemelidir. Bu hipotez yapısal geçerlilik çalışmalarında testin neyi ölçtüğü kadar neyi ölçmediğininde gösterilmesi ihtiyacını örneklemektedir. Eğer bütün bu hipotezler desteklenirse OPQ'nun yapı geçerliliğinin "Sözel Kötümserlik" olarak adlandırılan bir kişilik belirtisinin ölçümü olarak sağlandığını söylemek mümkün ve mantıklıdır. Test geçerliliğini göstermede daha güçlü ve dolaysız bir yaklaşımda, beşinci hipotezi oluşturabilir.

5. Sözel kötümserlik psikolojik özelliğine yüksek derecede sahip olanlar, düşük derecede sahip olanlara göre OPQ'dan daha yüksek puan elde ederler.

Yapı geçerliliği tek bir ölçütün oluşturulmasının zor olduğu testlerde, test geçerliliğini göstermede en güçlü yöntemdir. Tek bir sonuçla ilgilenmek yerine aynı anda bir grup sonuçla ilgilenmek zorundayız.

Yapı geçerliliğinin arttırılmasında giderilmesi gereken önemli bir problem, yapı geçerliliği sonuçlarının yorumlanmasında kullanılan öznel faktörlerle ilgilidir. Pratikte sıklıkla karşılaşıldığı gibi, çok açık olmadığı durumlarda yapı geçerliliği sonuçlarının yorumlanması, daha çok testi geliştiren araştırmacının yorumlama yeteneğine bağlıdır.

ÖZET

Testlerin geçerliliğini göstermek için farklı teknikler incelendi. Bunlardan bazıları diğerlerinden belirgin bir biçimde farklıdır. Yapı geçerliliği, geçerli bir testi, ölçmeyi, amaçladığını ölçmek olarak tanımlamamız ile yakından bağlantılıdır. Bu nedenle özellikle testlerin psikolojik bilgiyi arttırmak için kullandığı durumlarda geçerliliğin en önemli boyutunu oluşturmaktadır. Diğer taraftan ayırt edici geçerlilik, belirli hedefler için bir testin geçerliliğini göstermeyi amaçlamaktadır. Bu, geçerlilik kavramının oldukça farklı bir şekilde ele alınışı olup, kullanışlılık (utility) kavramına yaklaşmış halidir ve testlerin pratik uygulamalarında geçerliliğin bu boyutu oldukça önemlidir.

Yukarıda sürdürülen tartışmalardan anlaşılması gereken, bir testin geçerliliğini gösterecek bir tek yolun olamayacağıdır. Geçerliliği tam olarak değerlendirebilmek için bir bulgu setinin dikkate alınması gerekmektedir. Birçok testin (toplam sayı içinde çok küçük bir oranda olsa bile) yapı geçerliliğinde olduğu gibi, hem kavramsal hem de pratik amaçlar için yüksek derecede geçerliliğe sahip olduğu gösterilebilir. Ayrıca test yapımındaki mantıksal yöntemler yoluyla testin geçerliliği hemen hemen garanti edilebilir.

Ayırtetme Gücü

İyi bir testin bir diğer özelliği ayırtetme gücüdür. Gerçekten de iyi bir puan dağılımına ulaşmak için, yüksek ayırtetme gücü test yapımcılarının ulaşmaya çalıştıkları amaçlardan birisidir. Bütün bireylerin aynı puanı elde ettikleri psikolojik bir testin değerini düşünecek olursak, ayırtetme gücünün önemi kendiliğinden ortaya çıkar. Testin dikkatli oluşturulmasıyla iyi bir ayırtetme gücü elde etmek mümkündür. Böylece testler diğer değerlendirme araçlarına karşı belirgin bir üstünlük kazanırlar. Genel olarak derecelendirmeler veya işaretlemelerin yaklaşık dokuz kategoride yapılabildiği bulunmuştur (Vernon 1950). Buna bağlı olarak dereceleme ölçekleri ender olarak dokuz kategoriden fazlasını içerir. Bunun anlamı, ele alınan konuların en iyi biçimde dokuz grupta toplanabildiğidir. Ayırtetme gücü Ferguson'un deltası ile ölçülür ve en yüksek değer puanların dikdörtgen olarak dağıldığı durumlarda elde edilir.

Geçerlilik, güvenirlilik ve ayırtetme gücü konusunu bitirmeden önce yukarıdaki bütün tartışmalarda vurgulanan ölçme modelini özetlemek ge-

rekir. Model hakkında az bi r bilgiye sahip olmak bile, test geliştirme sürecini anlamaya yardım edecek ve test geliştirmede kullanılan yöntem ve hesaplamalar için mantıklı bir istatistiksel temel sağlayacaktır.

"Ölçmede Hata" Klasik Test Teorisi

Ölçmede hata teorisi, klasik test teorisi olarak isimlendirilir. Çünkü bu teori, testlerin incelenmesinin başlangıcından beri psikolojik test geliştiren araştırmacılar tarafından kullanılmış en temel varsayımlardan geliştirilmiştir. Son dönemlerde, daha karmaşık ve üst düzey modeller geliştirilmiş olmasına rağmen klasik teorinin temel ilkeleri hâlâ kullanılmaktadır. Bu ilkeler, özellikle test geliştirme uygulamaları için değerlidir.

Gerçek Puan

Bu teoride herhangi bir psikolojik özellik için her bireyin gerçek bir puana sahip olduğu kabul edilir. Herhangi bir uygulamada bireyin o testten aldığı puan rastgele hata gözönüne alındığında, o bireyin gerçek puanından farklıdır. Bireye birçok kez aynı testi uygularsak, puanlar bireyin gerçek puanı etrafında dağılır. Normal olduğu kabul edilen bu dağılımın ortalaması gerçek puana yaklaşıktır.

Ölçmenin Standart Hatası

Gerçek puan, ölçmenin standart hatasının temelini oluşturur. Bireyin testten elde ettiği puanlarda büyük bir dağılım -değişkenlik- buluyorsak bu durumda kayda değer derecede ölçme hatası var demektir. Bu hata dağılımının standart sapması aslında bir hatanın belirleyicisidir. Gerçekte bütün bireyler için hatanın aynı olduğunu kabul etmek mantıklı olduğundan hataların standart sapması, ölçmenin standart sapması olur. Test-tekrar test güvenirliliği bir testin iki farklı uygulamasından elde edilen puanlar arasındaki korelasyon olduğundan, bu modelde test-tekrar test güvenirliliği arttıkça ölçmenin standart hatası küçülür. Ölçmenin standart hatası şu formül ile gösterilir.

$$Se = Sx \sqrt{1 - r_{x_1x_2}} \qquad (1.1)$$

Sx = Testin standart sapması

$r_{x_1x_2}$ = Test-tekrar test güvenirlilik katsayısı

Test Maddelerinin Evreni

Klasik hata teorisi herhangi bir testin ölçülmek istenen psikolojik özellikle ilişkili olan madde evreninden elde edilen, rastgele seçilmiş madde örneklemlerinden oluştuğunu kabul eder. Yani saplantı (obsession) özellikleri ile ilgili bir test geliştiriyorsak, maddelerimizin bütün saplantı özelliklerini içeren maddeler arasından rastgele seçilmiş bir örneklem olduğu kabul edilir. Bu madde evreni hipotetiktir, kavramsaldır.

Birçok durumda, maddeler rastgele seçilemez. Fakat, Nunnanlly'nin (1978) de belirttiği gibi, test geliştiricileri aynı etkiye sahip maddelerle farklılaşmayı amaçlamaktadırlar. Bu durumda, maddeler madde evrenini yansıtamadığı oranda test hatalı olacaktır.

Gerçek Puanın Maddelerin Evreni İle İlişkisi

Bu modelde bir bireyin gerçek puanı mümkün olan bütün maddelerin verilebildiği durumda bireyin alacağı puandır. Öyle ise testlerin hatası, elimizdeki madde örnekleminin maddelerin evrenini kapsama derecesini yansıtır. Bu modelin, testin uygulandığı bireyin psikolojik durumu, oda ısısı, testi uygulayan kişinin yeterliliği gibi ölçme hatasına dahil edilebilecek diğer faktörleri dışarıda bıraktığını belirtmek gerekir.

Klasik Test Modelin İstatistiksel Temelleri

Klasik modelin istatistiksel temelleri Nunnanlly (1978) tarafından oluşturulmuştur. Daha öncede belirttiğimiz gibi gerçek puan bireyin hipotetik bir madde evrenindeki puanıdır. Bu maddelerin evreni, maddelerin birbirleriyle korelasyonlarının bir korelasyon matriksini oluşturur. Bu matrikste, maddelerin birbirleri ile ortalama korelasyonu r_{ij}, maddeler arasındaki ortak bir esasın (özelliğin) derecesini gösterir. Örneğin birbirleriyle ilişkisi olmayan farklı testlerden alınan maddelerin birbirleriyle ortalama korelasyonu sıfır olur. Bu da oldukça doğru bir biçimde bu maddeler arasında ortak bir esasın olmadığını gösterir. Benzer olarak r_{ij}'nin etrafındaki korelasyon katsayılarının varyansı, hangi maddelerin paylaşılan ortak esastan ne derece ayrıldığını gösterir. Modelde, bütün maddelerin ortak esasın eşit bir miktarına sahip olduğu kabul edilir. Bu da, her maddenin diğer maddelerle olan korelasyon katsayısının ortalamasının bütün maddeler için aynı olduğu anlamına gelir. Bu, modelin temel varsayımıdır.

Klasik modelden, bir maddenin gerçek puanla olan korelasyon katsayısının,onun bütün diğer maddelerle olan korelasyonunun ortalamasının karekökûne eşit olduğu gösterilebilir.

$$r_{it} = \sqrt{\bar{r}_{ij}}$$

(1.2)

Bu duruma tam olarak maddelerin sayısı sonsuza yaklaştığı zaman ulaşılabilir fakat sadece 100 madde kullanıldığında bile korelasyon katsayıları üzerinde çok küçük bir etkisi vardır.

Test geliştiricisinin bakış açısından formül (1.2) büyük öneme sahiptir. Çünkü, büyük bir madde havuzu oluşturur ve bu havuzdan diğer maddelerle ortalama korelasyon katsayısının karekökü yüksek maddeler seçerse, tanımı gereğince geliştirdiği test gerçek puanla yüksek bir korelasyon katsayısı vermeli ve böylecede yüksek güvenirlilikte ve ölçme hatasından arınık olmalıdır. Formül(1.2) bir madde havuzundan madde seçiminin istatistiksel temelidir. Bu, ulaşılamayan maddelerin yapay olarak ilişki gösterdiği hız testlerinde uygulanamaz.

Maddelerin maddelerle ilişkilendirilmesi durumu aynı değişkenin paralel testlerine de uygulanabilir. Herbir test madde evreninden rastgele seçilmiş bir maddeler örneklemi olarak kabul edilir. Böyle rastgele seçilmiş maddelerin ortalamaları ve varyansları gerçek puanlardan yalnızca şansa dayalı olarak farklıdır. Ele alınan bütün eşitliklerde maddeler için olan standart puanlar testlerin standart puanları ile yer değiştirmelidir.

Formül (1.2) $r_{it} = \sqrt{\bar{r}_{ij}}$ olarak yazılabilir.

r_{it} = Test I'deki puanlar ile gerçek puanların korelasyonudur

$\bar{r}_{ij}$ = Test I'in evrendeki tüm testler ile korelasyonunun ortalamasıdır.

Güvenirlilik Katsayısı

Bir testin ya da maddenin evrendeki diğer tüm testlerle ya da maddelerle olan ortalama korelasyonu güvenirlilik katsayısıdır. Bu katsayının kare kökü test yada maddenin gerçek puanla olan korelasyonudur (Formül 1.2). Fakat bu güvenirlilik $\bar{r}_{ij}$uygulamada bilinemez çünkü, madde ya da testlerin sayısı sonsuz değildir ve testler rastgele (random) paralel değildirler. Bu da, bir testin güvenirliliğinin sadece kestirebileceği anlamına gelir.(r_{ii})

Pratikte güvenirlilik katsayısı bir testin diğer bir testle korelasyonunu bağlı olduğundan, bu kestirim çok doğru ve tam olmayabilir. Bu da, çok önemli olan testin ya da maddenin gerçek puanla olan korelasyonunun da doğru olmayan bir kestirim olabileceği anlamına gelir.

Hatalı Puanlar (Gözlenebilen Puanlar)

Puanlar, herhangi bir testteki gerçek puanlar ve ölçme hatalarının birleşiminden oluşur. Bir test yada madde için pratikte elde ettiğimiz güvenirlilik katsayısı r_{ii}, $\bar{r}_{ii}$'ye yaklaşır. Eğer $r_{ii} = \bar{r}_{ij}$ kabul edersek r_{it} (gerçek ve gözlenen puanların korelasyonu) = r_{ii} olur. Böylece r_{it} kestirilebilir. Buradan gözlenen puanlardan gerçek standart puanların kestirilmesi aşağıdaki formül ile elde edilir.

(1.3)

$$Z't = r_{it} Z_i = \sqrt{r_{ii}} Z_i$$

Z't = kestirilmiş gerçek standart puanlar

z_i = gözlenen ölçümdeki standart puanlar

r_{it} = gerçek ve gözlenen puanların korelasyonu

r_{ii} = değişken I'in güvenilirliği

Bir değişkenin varyansına eşit olan korelasyon katsayısının karesi, diğer değişken açısından açıklanabildiğinden r_{it}^2, gerçek puan varyansının yüzdesi, hatalı ölçümle açıklanabilir. Fakat $r_{it} = r_{ii}$ olduğundan güvenirliliğin karesi hatalı ölçümdeki gerçek puan varyansının yüzdesine eşit olur.

Gerçektende Nunnanlly (1978)'nin gösterdiği gibi eğer test puanları (standart puan değil) sapma ya da ham puanlar ise;

(1.4)

$$r_{ii} = \frac{\sigma_{\pm}^2}{\sigma_i^2}$$

σ_i^2 = I. değişkenin varyansı

σ_t^2 = I. değişkenin gerçek puanlarla açıklanan varyansı

r_{ii} = güvenirlilik katsayısı

r_{ii} ve σ_i^2 kolaylıkla hesaplandığı için bu σ_t^2'nin kolay bir kestirimidir. Buradan, verilen bu klasik test modelinde, güvenirliliğin oldukça önemli olduğu çıkarılabilir.

Test Homojenliği ve Güvenirlilik

Bir testin güvenirliliği maddeler arasındaki ortalama korelasyon ile ilişkilidir ve bu da o testin homojenliğidir. Fakat madde korelasyonları

tam olarak aynı olmadığından bunların ortalama etrafında bir dağılıma sahip olması gerekir. Ölçmenin klasik modelinde bu dağılımın normal olduğu kabul edilir. Bu kabulden yola çıkarak bütün madde evrenindeki maddeler arası korelasyon ortalamasının kestirimindeki standart hatanın hesaplanmasıyla, güvenirlilik katsayısının doğruluk derecesini kestirmek mümkündür.

(1.5)

$$\sigma_{\bar{r}ij} = \frac{\sigma_{rij}}{\sqrt{1/2\, k\,(k\text{-}ı) - 1}}$$

$\sigma_{\bar{r}ij}$ = evrende kirij'nin kestirimindeki standart hata

σ_{rij} = testteki madde korelasyonlarının standart sapması

k = testdeki madde sayısı.

Formül (1.5), kestirimdeki hatanın madde korelasyonlaının standart sapmasını k tane madde arasında mümkün olan korelasyon kareköküne bölünmesi yoluyla elde edilebileceğini gösterir. -1doğru serbestlik derecesini verir. Formül (1.5'e göre)

a) Kestirimin standart hatası arttıkça kendi aralarındaki korelasyonlar daha çok farklılaşır

b) k arttıkça stardart hata azalır, bu da, madde sayısı arttıkça kestirimin güvenirlilik katsayısının doğruluğunun artması demektir. Bu formül, güvenirliğin test homojenliği ve test uzunluğu ile arttğını gösterir.

Sonuç, yarı test güvenirliliğini hesaplamada kullanılan Spearman-Brown formüldür.

(1.6)

$$r_{kk} = \frac{k\bar{r}ij}{1+(k-1)\,\bar{r}ij}$$

rkk = testin güvenirliliği

k = madde sayısı

$\bar{r}ij$ = ortalama madde inter korelasyonu

Spearman-Brown formülü test yapımında oldukça kullanışlıdır. Üç değişik madde grubumuz olduğunu kabul edelim. a) 10 madde, b) 20 madde c) 30 madde. Maddeler arasındaki ortalama korelasyon da 0.20 olsun:

$$\text{a grubu için} \quad r_{kk} = \frac{10x0.20}{1+(9x0.20)} = 0.667$$

$$\text{b grubu için} \quad r_{kk} = \frac{20x0.20}{1+(19x0.20)} = 0.080$$

$$\text{c grubu için} \quad r_{kk} = \frac{30x0.20}{1+(29x0.20)} = 0.959$$

r_{kk} testin güvenirliliğidir ve bunun karekökü bize maddelerin gerçek puanla olan korelasyonunun kestirimini verir. 10 maddelik bir test bile, kabul edilebilir bir güvenirlilik katsayısı verirken 30 madde ile oldukça yüksek bir değere ulaşılır. Bu değerler interkorelasyonları düşük olan (0.20) maddeler ile elde edildi. Ortalama korelasyonun 0.40'dan daha yüksek olduğu ve dolayısıyla daha homojen bir testde;

D grubu= 30 madde ve $\bar{r}ij$ =0.40

$$r_{kk} = \frac{30x0.40}{1+(39x0.40)} = \frac{12}{13} = 0.923 \text{ olarak bulunur,}$$

Büyük bir homojen madde havuzu oluşturabilen bir test geliştiricisi, güvenilir bir test oluşturabilir. Ayrıca 30 maddeyi 15'er maddelik iki paralel forma bölersek bu iki testinde oldukça iyi düzeyde güvenilir olacağını belirtmek gerekir. Öyleyse r_{kk} bize k maddelik bir testin aynı evreden bir diğer k maddelik test ile beklenen korelasyonunu verir. r_{kk} test maddelerinin interkorrelasyonlarından hesaplanan bir güvenirlilik katsayıdır.

Spearman-Brown (1.6) testin yarı test güvenirliliğin hesaplanmasında kullanılır (yarılar arasındaki korelasyonun testin uzunluğuna göre düzeltildiği durumlarda). Bu durumda testin her iki yarısı, evrenden seçilmiş bir örneklem olarak kabul edilir. Bu da özel durumlar için (k=2) formülün basitleştirilmesine olanak sağlar. Yarı test güvenirliliğinde kullanılan Spearman-Brown formülü aşağıdaki gibidir.

$$r_{kk} = \frac{2r_{12}}{1+r_{12}}$$

r_{12} = testin iki yarısı arasındaki korelasyon

Gerçekte temel formül (1.6) testin uzunluğunu, sayısını gözönüne alınmaz.

GÜVENİRLİLİK VE MADDE ÖRNEKLEMELERİ

Bu ölçme hatası modelinde, testlerin güvenirliliğini hesaplama yöntemleri kendi istatiksel temellerine sahiptir. Spearman-Brown formülü (1.6) bir testin güvenirliliğini hesaplamak için kullanılabilir. Fakat korelasyon matriksinin hesaplanması uzun sürdüğü için farklı görünmesine rağmen temelde aynı olan diğer yöntemler geliştirilmiştir.

ALPHA KATSAYISI

Cronbach (1971) ve (1978) alpha katsayısını test güvenirliğinin en önemli göstergesi olarak kabul etmektedir. Alpha katsayısı formülü ölçme hatasının klasik modelinden çıkarılmıştır ve daha basittir. Alpha katsayısı bir testin madde evreninden aynı uzunuktaki bir başka testle olan korelasyonun kestirimini gösterir.

(1.7)

$$\text{Alpha katsayısı} = \frac{k}{k-1}\left[1 - \frac{\sum \sigma i^2}{\sigma y^2}\right]$$

k = madde sayısı

$\sum \sigma i^2$ = madde varyanslarının toplamı

σy^2 = test varyansı

Alpha katsayısının karekökü testin gerçek puanla olan kestirilmiş korelasyondur.

KR-20 Kuder-Richardson, ikili puanlanmış (dichotomous) maddeler için alpha katsayısının özel bir durumudur.

$$r_{kk} = \frac{k}{k-1}\left(1 \frac{\sum PQ}{\sigma y^2}\right)$$

(1.8)

P=doğru cevap verenlerin oranı

Q=1-P ve σy^2 = testin varyansı

KR-20 alpha katsayısının taşıdığı özelliklere sahiptir ve hesaplanması kolaydır. İkili puanlanmış durumlarda PQ, σ^2'nin eşiti olur.

Alpha formülünden güvenilir bir testin güvenirliliği düşük bir teste oranla daha büyük bir varyansa ve dolayısıyla daha yüksek ayırt ediciliğe sahip olduğu çıkarılabilir.

Formül (1.5) den yapılan bu çıkarımlar test geliştiricileri için oldukça kullanışlı ve önemlidir. Öncelikle madde korelasyon kestiriminin standart hatasının ne anlama geldiğini hatırlatmak gerekir.

Bütün örneklemdeki ortalama korelasyonların %68'i ortalamanın bir stardart sapma altı ve üstü arasında, %95'i ortalamının iki stardart sapma altı ve üstü arasında bulunur. Bir testteki korelasyonların stardart sapmasının 0.15 olduğunu kabul ederek ve 10,20,30 maddelik testlerde formül (1.5)'i uygularsak aşağıdaki standart hataları buluruz.

10 maddelik test için 0.02

20 maddelik test için 0.01

30 maddelik test için 0.007

Bu sonuçlardan görülüyorki 10 maddelik bir test içi bile kestirilmiş güvenirliliğin doğruluğu oldukça yüksektir. Bunun nedeni de formül 1.5'in paydasının madde sayısı arttıkça hızla büyümesidir.

Bu kestirilmiş güvenirliliğin kesinliği test geliştiren araştırmacılar yönünden en büyük teşviktir. Bu, uygulamada güvenirliğin kestiriminde madde seçimindeki rastgele hatalardan kaynaklanan çok az bir hata olduğu anlamına gelir. Bir diğer önemli çıkarım da Nunlly'nin (1978) belirttiği gibi, görünürde paralel olan testler kendi aralarında düşük korelasyon gösterdikleri zaman, bu durum madde seçimindeki rastgele hatalara yüklenemez. Ya maddeler farklı madde evrenlerini temsil ediyorlardır (farklı değişkenleri ölçüyorlardır) ya da deneklerden kaynaklanan örneklem hatası vardır.

Böylece formül 1.5 test geliştiricilerine rastgele hataların, test yapımında kullanılan analizlere kolaylıkla zarar veremiyeceği yönünde güvence sağlar. Çok az bir madde sayısıyla bile güvenirlilik kestirimi yeterince kesin olabilir.

Klasik modelin gücü ve önemli, test yapımı için birçok kullanışlı çıkarımın yapılabiliyor olmasındadır.

GÜVENİRLİLİK VE TEST UZUNLUĞU

Güvenirlilik test uzunluğu ile birlikte artar. Gerçek puanlar madde evrenindeki puanlar olarak tanımlandığından testteki madde sayısı arttıkça gerçek puanla olan korelasyon da artar.

Test geliştirtiren araştırmacı için önemli olan madde sayısındaki artış ile birlikte güvenirlilikteki artış oranıdır. Çok sayıda geçerli madde geliştirmek zordur. Bunun sonucu olarak, diyelimki 25 madde ile güvenirliliğin yüksek olduğunu gösterebilirsek bu, mantıklı ve ulaşılabilir bir hedef olur.

Hata ölçümünde klasik modeli kullanarak, gözlenen puanlardan gerçek puanı kestirmek mükündür. Fakat bu test süreci ile ilgili değildir ve çok az bir pratik değere sahiptir.

Modelden çıkarılabilecek bir diğer yararlı istatistik, ölçmenin stardart hatasıdır.Bu, çok sayıda random paralel testleri alan bir kişi için puanların beklenen standart sapmalarıdır. Ölçmenin standart hatası gözlenen puanların güven aralığını belirlemek için kullanılır. Bu bölgeler, gözlenen puan etrafında değil gerçek puan etrafında simetri oluşturur. Fakat bu uygulamada genellikle dikkate alınmaz.

$$Se = Sx\sqrt{1 - rxt}$$

$$Se = Sx\sqrt{1 - rxx}$$

x = gözlenen puanlar

t = gerçek puan

rxx = güvenilirlik

Öyleyse ölçmenin stardart hatası gözlenen puanlardan kestirilmiş gerçek puanların standart hatasıdır.

STANDARDİZASYON VE NORMLAR

İyi geliştirilmiş psikolojik testlerin bir başka özelliği de iyi hazırlanmış normlardır. Normlar, çok iyi tanımlanmış örneklemelerden elde edilmiş puan gruplarıdır. Bu puanları elde etme yöntemleri ve geliştirilmesi test standardizasyonunu oluşturur.

Normlar tesleri kullanan kişilere, uygulama sonucunda elde edilen puanların kullanıldığı ve normların çok az bilgi eklediği araştırma amaçlı kullanımlardan çok, testlerin pratikteki uygulamalarında büyük değer taşır.

Örneklem grubunun uygun ve yeterli büyüklükte olması gereklidir. Aksi takdirde test normları kullanışsız olmanın ötesinde yanıltıcı olur. Eğer stardardizasyon uygun bir şekilde yapılmışsa psikolojik testler bize standardize edilmemiş yöntemlerin sağlamadığı bir karşılaştırma imkanı verir.

Testlerin dışında kalan değerlendirme yöntemlerinin bir çoğunun standardize edilemediği gözönünde bulundurulursa, standardizasyon, psikometrik testlerin önemli bir özelliği olarak ortaya çıkmaktadır.