

İKİ ELEŞTİREL AKIL YÜRÜTME GÜCÜ ÖLÇEĞİNİN GEÇERLİK VE GÜVENİRLİĞİNİN BİR PSİKOLOGLAR PANELİ TARAFINDAN DEĞERLENDİRİLMESİ*

Richard, B. MODJESKI

William, B. MICHAEL

**Çev: Nükhet ÇIKRIKÇI

ÖZET

Bu çalışma iki eleştirel akıl yürütme gücü ölçeğinin a) Cornell Eleştirel Akıl Yürütme Gücü testi Düzey X'le b) Watson-Glaser Akıl Yürütme Gücü Ölçeği-Form YM'nin geçerlik, güvenirlik ve ölçme hatalarının bir psikologlar paneli tarafından, değerlendirilmesi çalışmasıdır. Bu değerlendirmede ölçütler takımı olarak, "Eğitimsel ve Psikolojik Testlere İlişkin Standartlar" adlı yayının 1974 baskısındaki 10 temel geçerlik, 5 temel güvenirlik ve ölçme hatası standardı ele alınmıştır.

Kullanılan 15 standartın herbiri için iki eleştirel akıl yürütme testi ile ilgili olarak, her bir panel üyesi ölçeğin, o standartı karşılama derecesine göre "standartı tümüyle karşılıyor belli ölçüde karşılıyor, en alt düzeyde karşılıyor ve hiç karşılmıyor" şeklinde bir tepkide bulunmuştur. Bu düzeylerden herbirine 3, 2, 4 veya 1 gibi puanlar verilmiştir; ele alınan herhangi bir standart o ölçme aracı için uygulanabilir değilse o standart için herhangi bir puan verilmemiştir.

Genel olarak, her iki ölçme aracının standartları karşılama düzeyleri "en alt düzeyde karşılıyor" ile "belli düzeyde karşılıyor" kategorileri arasında yer almıştır. Bu derecelemelere ait ortalamalar arasında istatistiksel açıdan manidar bir fark çıkmamıştır.

Bu testlerin yazarlarına araçlarının geçerlik ve güvenirliklerini artırmak için bir revizyon çalışmasının gerekli olduğunun düşünülmesi önerilmektedir.

* Educational and Psychological Measurement, 1983, (43) 1187-1197.

** Arş. Gör., A.Ü. Eğitim Bilimleri Fak., E.P.H. Bölümü.

Eleştirel akıl yürütmenin kavramsal yapısına ilişkin merkezi süreçlerin temsilinde oldukça farklı yönelimleri yansıtan iki ölçek yaygın olarak kullanılmaktadır: (a) Cornell Eleştirel akıl Yürütme Testi-Düzey X ve Y (bir el kitabıyla birlikte) (Ennis ve Millman, 1971, a.b) ve (Ennis, Millman ve Tomko, 1979); Watson-Glaser Eleştirel Akıl Gücü Ölçeği Form YM ve ZM (Watson-Glaser, 1964 c). Heride gönderme yapılmasını kolaylaştırmak için bunlar arasirasıyla Cornell ve Watson-Glaser olarak adlandırılmıştır.

Bu çalışmanın amacı, sözkonusu terslerden herbirinin hem geçerlik, güvenilirlik ve hemde ölçme hataları bakımından bir psikologlar grubuna yapılan incelemesini sunmaktır. Bu dereceleme ölçekleri, el kitaplarında verilen bilgilere göre "Eğitimsel ve Psikolojik Testlere İlişkin Standartlar"da "Gerekli" kategorisine giren geçerlik, güvenilirlik ve ölçme hatası standartları yada ölçütlerini ne derece karşıladıkları bakımından yapılmıştır. İnceleme, Cornell'in Düzey-X'i ve Watson-Glaser'in YM formu ile sınırlandırılmıştır. Böyle bir değerlendirme, herbir ölçeğin yeniden düzenlenmesi veya geliştirilmesi ihtiyacı duyulduğunda seçilen psikometrik özelliklerine ilişkin bilgi verilmesini sağlayacaktır. Ayrıca, eleştirel akıl yürütme ile ilgili yapı ve alt yapılarının işevuruk tanımlarını açıklığa kavuşturmada, test geliştirme çabaları ve gelecekteki araştırmalar için yeni yönelimlerde çok yakın olabilir.

YÖNTEM

Örneklem

Örneklemi oluşturan uzman grubu (Doktora düzeyinde), bir dergide eleştirel akıl yürütme konusunda en az bir makalesi yayımlanmış 12 psikologtan oluşmuştur. Gruptaki üyelerin herbiri A.B.D.'deki üniversitelerin birinde akademisyen olarak çalışmaktaydı. Gruptaki üyelerden biri Cornell Eleştirel Akıl Yürütme Testi İnceleme ölçeğine ilişkin soruları cevaplamadığından bu aracı 11 panel üyesi, WG-EAGÖ'ni ise 12 üye incelemiştir.

Dereceleme Ölçeği

Organizasyon ve Biçim: Ele alınan ölçeklerin belirlenen standartlara göre incelenmesini sağlamak üzere geliştirilen ilk 19 kişilik gruba gönderilen anket (12 kişinin yanıtladığı) üç bölümden oluşmuştur: (a) Eğitimsel geçmiş ve halihazırdaki istihdam durumu, mesleki yönelim-

lere ilişkin 6 temel madde içeren I. bölüm; Cornell'in ve Watson-Glaeser'in ön çalışmada ele alınan 10 temel geçerlik ve 5 temel güvenilirlik ölçme hatası standartını temsil edici, araçların bu standartları karşılama düzeyini ölçeklemeye izin veren 15 cümleden oluşan 2. ve 3. bölümlerdir. Panelistlere bunları ölçeklemede 5 seçenek sunulmuştur. Bunlar;

- A- Standartı tümüyle karşılıyor
- B- Standartı belli ölçüde karşılıyor
- C- Standartı en alt düzeyde karşılıyor
- D- Standartı hiç karşılamıyor
- E- Standart uygun değil

A, B, C, D, veya E seçeneklerine verilen cevaplar her bir standartla bağlantılı olarak 4, 3, 2, 1 ve 0 şeklinde ağırlandırılarak puanlanmıştır. Beşinci seçenek ele alınan standartın sözkonusu ölçme aracı için uygulanamaz olması durumunda kullanılan ve puan karşılığı sıfır olan bir seçenektir.

Her bir cümleye karşılık gelen sayısal değer, panel üyesinin görgül çalışmalardan çıkarılan bulgular ile testin el kitabından sağlanan bilgiler çerçevesinde, her testin bir standartı karşılama derecesi hakkındaki derecesini sayısal olarak ifade etme olanağı sağlamıştır.

Temel Standartlar

Bu çalışmada ele alınan testleri incelemede yararlanılan standartlar; geçerlik alanında temel olan 10 standart ile güvenilirlik ve ölçme hatası konusunda temel olan 5 standarttır. Tercih edilir, edilmez şeklinde sınıflandırılan diğer standartlar incelemede gözönüne alınmamıştır. E bölümünde ele alınan geçerliğe ilişkin standartlar, temel olarak sınıflandırıldıkları her durumda E₁, E₂, E₃, E₄, E₅, E₆, E₇, E₈, E₉, E₁₀, E₁₂, E₁₃ şeklinde kodlanmıştır. Buna karşılık güvenilirlik ve ölçme hatasıyla ilgili standartlar F bölümünde ele alınmış ve temel olarak sınıflandırıldıkları her durumda F₁, F₂, F₃, F₄, ve F₆ sembolleriyle gösterilmiştir.

Geçerlik Standartları

Genel İlkeler

E₁ Bir el kitabı veya araştırma raporu testin kullanılmasının önerildiği her çıkarsama türü için geçerliğe ilişkin kanıtlar içermelidir. Eğer geçerlik, öne sürülen bazı yorumlamalar için araştırılmamış ve incelenmemişse bu durum açıklığa kavuşturulmalıdır.

E₂ Bir test uygulayıcısı, geçerlik ve güvenilirlik iddialarını destekleyecek kanıtlar sunmakla yükümlüdür. Karara ulaşmada test puanlarının kullanımı, yeterli derecede kanıtla desteklenmiş olmalıdır.

Ölçüt Dayanıklı Geçerlik

E₃ Bütün ölçüt ölçüleri tam ve açık bir şekilde testin el kitabında belirtilmelidir. El kitabı veya araştırma raporu ölçütün uygunluğu konusundaki yorumlamalarda bulunmalıdır. Olabildiğince, performansın, ölçüt ölçüsünün yansıtmadığı yönlerine ve onu etkileyecek ilişkisiz etmenlere dikkat çekmelidir.

E₆ Bir geçerlik saptama çalışmasında, seçilen örneklem ve testin uygulandığı koşullar, testin kullanımı için önerilen koşullarla tutarlı ve okuyucunun kendi durumu ile ilişkisi konusunda karara varabilmesi için yeterince açıklanmalı ya da betimlenmelidir.

E₇ Bir geçerlik çalışması için verilerin toplanmasında, çalışmanın amaçları ile tutarlı işlemler izlenmelidir.

E₈ Birey hakkında yapılan tahmin ve yargılamalara ne kadar güvenle bakılabileceği konusunda okuyucuya yardımcı olacak "ölçüt dayanıklı geçerlik" verilerine ilişkin her türlü istatistiksel analize teslim el kitabında yer verilmelidir.

E₉ Bir test kullanıcısı testteki veya test maddelerindeki yanlışlık olasılıklarını incelemelidir. Testi kullanan kişi için, mümkün olduğunca etnik, cinsiyet ve diğer alt gruplar (kültür) açısından, ölçüt dayanıklı geçerlikte ortaya çıkması olası farklılıkların araştırılması ve bu gruplara ilişkin verilerin el kitabında bulunması, açıklanması gerekir.

E₁₀ Bir puanlama anahtarına sahip ölçme aracının maddelerinin seçimi veya testin (maddelerinin) ağırlıklandırılması, yalnızca bir tek örnekleme dayalı olarak yapıldığında, el kitabının bağımsız örneklemlerden elde edilen verilere dayalı olarak, bir veya daha fazla sayıdaki çapraz geçerlik (Cross-validation) katsayılarını içermesi gerekir. Geçerlik ifadeleri orijinal örnekleme bağılı olarak verilmemelidir.

Kapsam Geçerliği

E₁₂ Eğer testteki performans, evrendeki performansı temsil eden bir örneklem olarak yorumlanıyorsa; testin el kitabı, evreni ve evrenden alınan örnekleme seçmede takip edilmiş olan işlemleri açık bir biçimde tanımlamalıdır.

Yapı GeçerliĐi

E₁₃ EĐer yazar, bir testten alınan puanları teorik bir deĐişkeni (yetenek, tutum v.b. psikolojik özellikleri) ölçüyor olarak yorumlamayı önerirse, önerdiĐi yoruma tam olarak yer verilmelidir. Yazarın öne sürdüĐü kuramsal yapı, diĐer kuramlar temelinde ortaya çıkan yorumlamalardan ayırddedilebilmelidir.

Güvenirlik ve Ölçme Hatası Standartları*Genel İlkeler*

F₁ Testin el kitabı veya araştırma raporu testin amaçlanan kullanımını açısından puanların yeterli olup olmadığı kullanıcının yargılamasına olanak veren güvenilirlik hesaplamalarını içermelidir.

F₂ Güvenirlik katsayılarını ya da ölçmenin standart hatasını belirlemede kullanılan işlem ve örneklemeler, test uygulayıcısının, rapor edilen verileri ilgilendiĐi birey veya gruplara uygulanabilir nitelikte olup olmadıklarına karar verebilmesine olanak sağlamalıdır.

F₃ Güvenirlik çalışmalarına dair raporlar, sırasıyla hata bileşenlerinin varyansları ölçmenin standart hatası ya da çarpım momentler korelasyon katsayıları cinsinden el kitabında ifade edilmelidir. Verilerin iyi bilinmeyen bir biçimde ifadelendirilmiş olmaı durumunda, gelişimlerine gündermeler yapılacak açık bir biçimde betimlenmesi gerekir.

Formların KarşılaştırılabilirliĐi

F₄ EĐer bir testin aynı cevaplayıcı grubu için kullanılmak üzere geliştirilmiş iki veya daha fazla formu varsa, formlardaki ortalama, varyans ve madde özellikleri hakkındaki bilgiler, bu formlardan alınan puanlar arasındaki korelasyon katsayılarıyla birlikte el kitabında raporlaştırılmalıdır. EĐer bu türden kanıtlar elde edilmemişse, testin el kitabı okuyucuyu, formların denkleĐini kabul etme konusunda uyarmalıdır.

Formların Uygulamalararası Sonuçlarının Karşılaştırılması

F₆ Testin el kitabı puanların ne ölçüde kararlı olduğunu yani, belirli bir zaman aralıĐıyla bir testin paralel formu uygulanmışsa, puanların ne kadar kararlılık gösterdiĐi belirtmelidir. Ayrıca tesin kullanılabilirliĐindeki böylesi herhangi bir varyasyonun etkisi de el kitabında açıklanmalıdır. İki uygulama arasında geçen sürenin uzunluĐu testin doğasına ve test puanlarının nasıl yorumlanacağına baĐlıdır.

İstatistiksel Analizler

Verilen bir standart için alınan cevap kategorilerine ilişkin olarak yapılan derecelemelerin sıklığı her ölçme aracı için bulunmuş ve raporlaştırılmıştır. Panel üyeleri tarafından tayin edilen derecelemelerin ortalaması, medyanı, ranjı ve standart kayması herbir standart için belirlenmiştir. Buna ek olarak hariki ölçme aracı için geçerlik, güvenilirlik standartlarını karşılama düzeyleri bakımından yapılan ortalama derecelerin ortalamaları bulunmuştur.

Ayrıca panel üyelerinin herbir geçerlik, güvenilirlik ve ölçme hatası standartı için tayin ettikleri ortalama dereceleme puanlarının ortalamaları arasındaki farkın manidarlığı yine her iki araç için test edildi. Son olarak panel üyelerinin 14 standart için tayin ettikleri derecelendirmelerin tümüne ilişkin kategori iç derecelemelerin hesabı toplam veya ortalama olarak elde edildi (Guilford, 1965, s. 300).

BULGULAR

Panel üyelerince yapılan derecelemelere ait veriler her iki ölçek için: a) yargıcılararası güvenilirlik tahminleri açısından b) çoğunluğu betimleyici olmak üzere, panel üyeleri tarafından iki aracın 10 geçerlik standartını karşılama derecesine ilişkin olarak derecelmeler açısından c) iki aracın beş güvenilirlik ve ölçme standartlarını karşılama derecesine ilişkin olarak yapılan derecelemelerin paralel betimsel istatistikleri aşağıda verilmiştir.

1. CET ve WG-EAGÖ için sırasıyla 11 ve 12 psikolog tarafından, ortalama derecelemelere dayanılarak hesaplanan "yargıcılararası güvenilirlik katsayısı (r_{kk})".84 ve .88 (— .001) olarak hesaplanmıştır.

2. Panel üyelerinin Cornell ve Watson-Glaser için 10 geçerlik standartını karşılama düzeylerine ilişkin olarak verdikleri derecelmeleri için aşağıdaki istatistiksel sonuçlar Tablo 1'de gösterilmiştir.

a) Beş dereceleme kategorisi (A, B, C, D veya F) üzerinden yapılan derecelmelerin dağılımındaki önemli ranjlar her iki ölçümde vardı ve her standarda ilişkin olarak hemen her kategori en az bir frekans alıyordu. Örneğin F2 standartı, her kategoriye en az bir defa, sözkonusu araç için konmuştu.

b) Eğer bir geçerlik standartı "olumlu (tercih edilebilir) derecelendirilmek için panel üyelerinden yüzde 50'sinden ya da daha fazlasından

ya Standardı tümüyle karşılar (kategori A) ya da "Standardı kısmen karşılar (kategori B)" derecelendirmesi âlmak "zorundadır." ölçütü konulursa Cornell için iki standardın (F_6 ve F_7), Watson- Glaser için beş standardın (E_1, E_2, E_6, E_7 ve E_{12}) ölçütü karşıladığı gözlenmiştir.

c) Her iki ölçme aracı da E_9 ve E_{10} standartları açısından, en az sıklıkla A ve B kategorilerinde yer almıştır.

d) 10 geçerlik standardına ilişkin olarak yapılan derecelemeler üstünden hesaplanan ortalama ve standart sapmalar Cornell için, 2.22 ve .42; Watson-Glaser için, 2.31 ve .48'dir.

Bu işlemde, bir derecelendiricinin (panel üyesi) 10 geçerlik standardı açısından, iki ölçme aracı için de yaptığı iki dereceleme takımının birbirinden bağımsız olduğu sayılışı vardır. Bu sayılışı, bir derecelendiricinin belli bir standarda karşı olumlu-olumsuz yanlıklarının bir cevap kurulumu (response set) yaratma olasılığı açısından sorgulanmıştır. Bu amaçla hesaplanan ortalamalar arasındaki .09'luk fark istatistiksel olarak manidar bulunmamıştır ($t(18) = .43, P < .05$).

3. Cornell ve Watson-Glaser güvenilirlik ve ölçme hatasına ilişkin beş standardı karşılama düzeylerine ilişkin olarak panel üyelerinin yaptıkları derecelemelerden çıkan ve aşağıda verilen istatistiksel sonuçlar Tablo 1'de özetlenmiştir.

a) Geçerlikteki gibi, güvenilirlikte, beş kategoride verilen cevapların sıklığına ilişkin yapılan derecelemeler, geçerlikte ki derecelemelerden daha yüksektir.

b) Eğer aynı standart sözkonusu araç için cevaplayıcıların yüzde50'si tarafından A ve B dereceleme kategorisine konmuşsa, verilen standart ifadesinin, güvenilirlik ve ölçme hatası standartına uygunluğunu temsil ettiği kabul edilmiştir. Bu bulgu, Cornell için F_2 ve F_3 standartlarında; Watson-Glaser için F_1, F_2, F_3 ve F_4 standartlarında sağlanmıştır.

c) Her iki ölçme aracı F_6 standardı açısından A kategorisinde bir derece almamış B kategorisinde de sadece bir dereceleme elde etmiştir.

d) Sözkonusu güvenilirlik ve ölçme hatası standartları açısından yapılan beş derecelemenin ortalama ve standart sapması sırasıyla, Cornell için, 2.46 ve .74; Watson-Glaser için, 2.74 ve .79 bulunmuştur.

Yine iki ölçme aracı için yapılan beşli iki takım dereceleme işleminin birbirinden bağımsız olduğu sayılışını test etmek için, bu derecelemelere ait ortalamalar arasındaki .28'lik fark istatistiksel olarak incelenmiş ve manidar bulunmamıştır ($t(18) = .42, P < .05$).

SONUÇLAR

Geçerlik ve güvenilirlik standartlarından yola çıkılarak bir grup uzmanın bu çalışmada yaptığı derecelemelerin yorumu şu şekildedir:

1. Geçerlik konusunda ileri sürülen genellemeler,

a) Geçerlikte, "temel" olarak sınıflandırılan standartlar açısından sözkonusu ölçme araçlarının bu standartları karşılama düzeylerindeki, panel üyesi psikologların algılamaları arasında bir değişim gözlenmiştir.

b) Her iki araç, nispeten geçerlik hesaplamalarında başvuru olan iki örneklem hakkında gerekli bilgiyi vermeleri bakımından (E_6 standardı) ve böylesi bir amaç için veri toplam koşulları gözönüne alındığında, oldukça yüksek puan düzeyine sahip cevap kategorilerinde derecelendirilmişlerdir.

Her iki ölçme aracı da olası test yanlılığı bağlamındaki değerlendirmeleri ve çapraz geçerlik çalışmalarındaki eksikliği karşılama bakımından oldukça yüksek derecelmeler elde ettiler.

Watson-Glaser, tek bir standart açısından (evreni yansıtıcı davranışları örneklemleyen maddelerin seçiminde takip edilen işlemlerin tanımlanması ile evrenin tanımlanması standardı) E_{12} Cornell'e göre daha üstün olarak görülüp, derecelendirilmiştir.

c) Her iki araç için 10 geçerlik standardına ilişkin olarak, panel üyelerinin tayin ettiği derecelerin ortalamaları arasında manidar bir fark gözlenmemesine rağmen, Watson-Glaser temel geçerlik standartları açısından Comell'e göre daha üstün bir araç olarak görülmüş, panel üyelerinin en az % 50'si tarafından bu araçlar A ve B kategorilerine (Watson-Glaser dört standart açısından, Cornell ise iki standart açısından) atanmıştır.

2. Güvenirlik ve Ölçme Hatasına İlişkin Olarak İleri Sürülen Genellemeler

a) Geçerlikte olduğu gibi, bunda da her iki ölçme aracının temel standartlardaki beş standardı karşılama düzeylerine ilişkin olarak panel üyelerinin derecelmeleri arasında bir değişim çıkmıştır.

b) Her iki araç da (i) "güvenirlik katsayılarının ve ölçme hatalarının belirlenmesinde kullanılan örneklemin seçimi ve takip edilen işlemlerin tanımlanmasında" (F_2 standardı); (ii) "Güvenirlikle ilgili verilerin raporlaştırılmasında kabul edilen yöntemlerin kullanılması bakımından" (F_3 standardı), tercih edilir araçlar olarak belirlenmişlerdir.

Cornell ve Watson-Glaser, her ikisi de "paralel formların kullanılması bağlamında, puanların kararlılığının raporlaştırılmasını içeren standartta ve bu paralel formların uygulamaları arasında geçen süre ile ilgili standartta" (F_6 standardı) negatif uçta derecelmeler elde ettiler.

Aşağıdaki standartlar açısından Watson-Glaser, Cornell'e göre üstünlük sergilemiştir:

Watson-Glaser, Cornell'e göre (i) ölçme hatası ve güvenilirlik ile ilgili araştırma raporları ve el kitabında bununla ilgili kanıtlar bulunması (F_1 standardı) (ii) testin iki veya daha fazla formunun birbirine denkliği konusundaki verilere sahip olması (F_4 standardı) açısından daha üstün bir araç olarak gözlenmiştir. (Cornell de eleştirel akıl yürütmenin karşılaştırılabilir düzeylerini yansıtacak iki formun düzenlenmesinin eksikliği, muhtemelen bu aracın tercih edilmeyen bir şekilde derecelenmesine yol açmıştır). (iii) Panel üyelerinin ölçme hatası ve güvenilirlik standartları açısından yaptıkları derecelemelere ait ortalama puanlar arasında manidar fark olmamasına rağmen, Watson-Glaser temel standartları karşılaması bakımından daha üstün bir araç olarak görülmüştür. Elde edilen derecelmelerin herhangi biri, panel üyelerinin en az % 50'si tarafından "standartı tümüyle karşılıyor" (A kategorisi) ve "standartı belli ölçüde karşılıyor" (B kategorisi) kategorisinde, sırasıyla, Watson-Glaser dört standartta, Cornell iki standartta derecelmeler elde etmiştir.

ÖNERİ

Eğer standartlardan sağlanan ölçütler, iki aracın değerlendirilmesi için uygunsa, güvenilirlik, geçerlik belirlenmesine ilişkin araştırma ve geliştirme çabaları ortaya çıkabilecektir. Gelecekte her iki aracın da gözden geçirilmesi düşünülmelidir.

Kaynaklar

- American Psychological Association (1974). Standards for educational and psychological tests. Washington, D.C.: Author.
- Ennis, R.H. and Millman, J. (1971 b). Cornell Critical Thinking Test: Level Z. Urbana, Illinois: University of Illinois, Illinois Rational Thinking Project.
- Ennis, R.H., Millman, J., and Tomko, T.N. (1979). Manual for Cornell Critical Thinking Test: Level X and Level Z (rev. ed.) Urbana, Illinois: University of Illinois, Illinois Rational Thinking Project.

Guilford, J.P. (1965). *Fundamental statistics in psychology and education* (4 th ed.). New York: Mc Graw-Hill.

Watson, G. and Glaser, E.M. (1964 a). *Watson-Glaser Critical Thinking Appraisal: Form YM*. New York: Harcourt, Brace and World.

Watson, G. and Glaser, E.M. (1964 b). *Watson-Glaser Critical Thinking Appraisal: Form Zm*. New York: Harcourt, Blace and World.

Watson, G. and Glaser, E.M. (1964 c). *Watson-Glaser Critical Thinking Appraisa Manual: Form YM and ZM*. New York: Harcourt, Brace and World.