BÖLÜM 2

GENİŞ ÖLÇEKLİ SINAVLARDA GEÇERLİK VE GÜVENİRLİK: PISA 2018 ÖRNEĞİ

Dr. Şule ÖTKEN[1]

DOI: https://dx.doi.org/10.5281/zenodo.8417063

[1] Dr. Şule ÖTKEN.Matematik Öğretmeni
Eğitimde Ölçme ve Değerlendirme Ana Bilim Dalı. sule.ayyildiz@hotmail.com

1.GİRİŞ

Sosyal bilim araştırmalarının önemli bir kısmı insan davranışının ölçülmesidir - yani insan davranışını gözlemlemek için ölçüm araçlarının kullanılmasıdır. İnsan davranışının ölçümü, gerçekliği ayırt etmek için yaygın olarak kabul gören pozitivist görüşe veya ampirik-analitik yaklaşıma aittir (Smallbone & Quinton, 2004). Ölçüm araçlarında bulunan soruların yapılan araştırmaların doğruluğunu ortaya çıkarabilecek özellikte, birbiri ile tutarlı, anlaşılır ve yeterli sayıda olması dikkat edilmesi gereken noktalardan bazılarıdır (Kalaycı, 2014, s. 403). Ölçeğin standartlaştırılarak uygun bilgiler üretme yeteneğine sahip olması için "güvenirlik" ve "geçerlik" olarak nitelendirilen iki temel niteliğe sahip olması gerekir (Ercan ve Kan, 2004, s. 211). Bu makalenin amacı, geçerlik ve güvenirlik hakkında alanyazına dayalı bilgiler sunarak geniş ölçekli sınavları bu kapsamda değerlendirmektir. Bununla birlikte Ekonomik İş Bilriği ve Kalkınma Örgütü (OECD) tarafından PISA sınavlarının da geçerlik ve güvenirlik konularında çerçevesinde değerlendirilerek önerilerde bulunmaktadır.

1.1.Geçerlik

Geçerlik, bir değişkenin ölçmesi gereken şeyi ölçüp ölçmediğiyle ilgilidir. Geleneksel olarak, psikologlar dört tür geçerlilik ayırt etmişlerdir: kapsam geçerliliği, ölçüt geçerliliği, yapı geçerliliği ve yakınsak ve ayırt edici geçerlilik. Her biri bir ölçümün bir kavrama karşılık gelip gelmediğini göstermeye çalışır, ancak bunu yapma araçları farklıdır (Bollen, 1989, s.185). Gulliksen (1959)'ye göre bir testin geçerliliği, testin bazı kriterlerle olan korelasyonudur (Ebel, 1961, s.640). Geçerlilik, ampirik kanıtların ve teorik gerekçelerin, test puanlarına veya diğer değerlendirme kriterlerine dayalı yorumların ve eylemlerin yeterliliğini ve uygunluğunu destekleme derecesine ilişkin genel bir değerlendirme yargısıdır (Messick, 1989, s.

Bir ölçme araç ve yönteminin geçerliği, ölçülmek istenen özelliği başka değişkenle karıştırmadan ölçebilmesidir. Başka bir deyişle bir ölçme aracının geçerliği aracın amacına hizmet etme derecesidir (Baykul ve Turgut, 2010, s. 131). Bir test, geçerli olmak için güvenilir olmak zorundadır. Öte yandan güvenilir olan bir test her zaman geçerli olmayabilir (Tekin, 1991, s. 43). Bir çalışmanın geçerliliği, sistematik hata derecesine bağlı olarak iki bileşene ayrılır. İç ve dış geçerlilik olarak bilinen iki tür geçerlilik vardır. İç geçerlilik, hastalığa maruz kalma ve değişkenler arasındaki ilişkiler de dahil olmak üzere ölçümlerdeki hata miktarına bağlıdır. İyi bir iç geçerlilik,

ölçümde hata bulunmadığı anlamına gelir ve çıkarımların yapılabileceğini gösterir. Dış geçerlilik, çalışmanın bulgularının örneklemin alındığı popülasyona genelleştirilmesi süreciyle ilgilidir. Hangi koşulların genelleme ile ilgili olduğunun anlaşılmasını gerektirir (Singh ve Masuku, 2012, s.91).

1.1.1.Kapsam geçerliği: Kapsam geçerliliği, bir kavramın etki alanının netleştirildiği ve analistin ölçümlerin etki alanını tam olarak temsil edip etmediğine karar verdiği nitel bir geçerlilik türüdür (Bollen, 1989, s.185). Test kullanıcısının, sınava giren kişinin test puanından testin kendisindekilere benzer daha geniş bir alana ilişkin çıkarım yapmak istediği durumlar için kapsam geçerliği kullanılır (Crocker ve Algina, 2008, s.217). Başka bir deyişle kapsam geçerliliği, bir değerlendirme aracının öğelerinin, belirli bir değerlendirme amacı için hedeflenen yapıyla ne ölçüde ilgili olduğu ve onu temsil ettiği derecesidir (Haynes, Richard ve Kubany, 1995, s.239). Kapsam geçerliği, testin ve testteki her maddenin amaca hizmet etme derecesidir. Kapsam geçerliği, bir testin toplam maddelerin ölçülecek davranışları ve konu içeriğini örnekleme derecesine ve testteki her maddenin ölçme istediği davranışı ne derece ölçtüğüne bağlı olarak değişir (Tekin, 1991, s. 45).

1.1.2.Yapı geçerliği: Yapı geçerliliği, içerik ve ölçütle ilgili kanıtlar da dahil olmak üzere test puanlarının yorumlanması veya anlamı üzerinde etkisi olan ve yapı geçerliliğinin bir parçası olarak kabul edilen her türlü kanıtın bütünleştirilmesine dayanır (Cronbach ve Meehl, 1955; Akt: Messick, 1994, s.3). Hipotez testinin yanı sıra yapı geçerliliği aynı zamanda yapısal geçerliliği ve kültürler arası geçerliliği de içerir. Yapısal geçerlilik, bir ölçüm aracına ait puanların, ölçülecek yapının boyutsallığının yeterli olma derecesini tahmin eder. Kültürlerarası geçerlilik, çevrilmiş veya kültürel olarak uyarlanmış bir sonuç aracındaki maddelerin performansının, sonuç araçlarının orijinal versiyonundaki maddenin performansının yeterli bir yansımasının derecesini tahmin eder Scholtes, vd., 2011, s.239).

1.1.3.Ölçüte dayalı geçerlik: Çoğu durumda test kullanıcısı, bir testle doğrudan ölçülemeyen bazı performans kriterlerine göre test puanlarından sınava giren kişinin davranışına ilişkin çıkarımlar yapmak ister. Bu tür kanıtlar, kritere bağlı geçerleme çalışmalarından elde edilir (Crocker ve Algina, 2008, s.224). Bir ölçme aracının geçerliği yeterli güvenirlik ve geçerlikte olduğu bilinen başka bir ölçme aracıyla olan korelasyonuna bakılarak hesaplanır. Korelasyon +1'e yakın ise hem yordayıcı hem ölçütün geçerliğinin yüksek olduğunu gösterir (Baykul ve Turgut, 2010). Eş zaman geçerliği ve yordama geçerliği olarak ikiye ayrılır. Eş zaman geçerliği,

katılımcıların geliştirilmek istenen testten aldıkları puanlarla aynı kişilerin aynı davranışı ölçen eski bir testten aldıkları puanlarla olan korelasyonuna bakılır. Yordama geçerliği, test puanı ile gelecekte ölçülecek davranış arasındaki ilişki incelenerek test sonuçlarının gelecekteki davranışı ne derece yordadığı incelenir (Büyüköztürk, 2009, s. 169).

1.1.4.Görünüş Geçerliği: Bir testin gerçekten ne ölçtüğü ile ilgili değil, o testin ölçmek istediği özelliği ölçüyor görünmesidir. Testteki her bir soru da aynı şekilde görünüş geçerliğine sahip olmak zorundadır (Tekin, 1991, s. 53).

1.2.Güvenirlik

Güvenilirlik, veri kalitesini değerlendirmek için temel ölçütlerden biridir. Güvenilirlik burada bir ölçüm veya hesaplamanın doğru olarak kabul edilebilme derecesi olarak tanımlanabilir (Wagemaker, 2020, s.11.) Başka bir ifadeyle güvenilirlik, gözlenen puanlardaki tutarlılık ve tutarsızlıkların ölçülmesini içerir (Brennan, 2011, s.2). Güvenilirlik herhangi bir ölçüm yöntemi için önemli olduğundan, yeni ölçümler geliştirildiğinde güvenilirlik araştırmaları yapılmalıdır. Bunlardan bazıları, iç tutarlılık yöntemi, alternatif formlar, Cronbach Alpha katsayısıdır (Nunnally ve Berstein, 1994, s.273).

Güvenilirlik, ölçümün tutarlılığıdır. Bir ölçümün temelinde tek bir gerçek puan olduğunu varsayan klasik test teorisinden ortaya çıkan bir kavramdır (Bollen, 1989, s.221). Güvenilirlik katsayılarının başlıca kullanımı, sonuçların tekrarlanabilirliğini bildirmektir. Her türlü geçerlilik için en azından bir miktar güvenilirlik gerekli olduğundan, güvenilirlik katsayısı bir aracın etkililiğinin bir göstergesidir (Nunnally ve Berstein, 1994, s.256).

Güvenilirlik, sınava girenlerin aynı testin tekrarlanan uygulamaları veya testin paralel formları üzerindeki göreceli performanslarının tutarlılığını ifade eder. Bireysel test performansındaki tutarsızlığın ana kaynağı tesadüfi ölçüm hatasıdır. Tesadüfi ölçüm hatalarının test performansını ne ölçüde etkilediğini belirlemeye çalışmak, test geliştiricilerinin ve test kullanıcılarının sorumluluğundadır. Klasik gerçek puan modeli, pratik güvenilirlik araştırmalarının geliştirilmesi için teorik bir çerçeve sağlar (Crocker ve Algina, 2008, s.119).

1.2.1.İç tutarlılık: Bir testin güvenirliği, testin eşdeğer yarıları arasındaki tutarlığa bakmak yerine bütün soruların birbiriyle tutarlığına bakmaktır. Söz konusu durumda KR-20(Kuder Richardson) formülü kullanılır

(Baykul ve Turgut, 2010, s.126). İç tutarlılık tahmini, ölçüm aracının tekrar tekrar uygulanmasının söz konusu olmaması nedeniyle diğer güvenilirlik tahminlerinden farklılık göstermektedir. Bireysel öğeler arasındaki karşılıklı ilişkinin derecesi tahmin edilerek ölçülür (Scholtes, Terwee ve Poolman, 2011, s. 237).

1.2.2.Test tekrar test yöntemi: Test-tekrar test güvenilirliği, bir testin bir ölçüm oturumundan diğerine zamansal istikrarını ifade eder. Prosedür, testi bir grup katılımcıya uygulamak ve daha sonra aynı testi aynı katılımcılara daha sonraki bir tarihte uygulamaktır. Farklı zamanlarda verilen aynı testlerdeki puanlar arasındaki korelasyon, test-tekrar test güvenilirliğini operasyonel olarak tanımlar (Drost, 2011, s.108). iki puan seti arasındaki ilişki Pearson korelasyon katsayısı ile hesaplanır. Testin zaman bağlı olarak ne derece kararlı ölçümler verdiği hakkında çıkarımlar yapılmasını sağlar (Büyüköztürk, 2009, s. 170).

1.2.3.İki yarı test güvenirliği: Test maddelerinin tek-çift, ilk yarı-son yarı gibi iki eş yarıya ayrılarak testin iki yarısı arasındaki ilişki hesaplanır. Söz konusu ilişkiyi hesaplarken Spearman Brown formlü kullanılır (Büyüköztürk, 2009, s. 170). Yarıya bölme yöntemi kullanıldığında, test geliştiricisi bir grup katılımcıya testin bir formunu uygular. Ancak testi puanlamadan önce, test geliştiricisi maddeleri her biri orijinal testin yarısı uzunluğunda iki alt teste böler (Crocker ve Algina, 2008, s.136).

1.2.4.Cronbach Alpha: Alfa katsayısı (α), icadından bu yana teorik ve pratik ölçüm araştırmalarında test puanlarının iç tutarlılık güvenilirliği için en yaygın kullanılan katsayı haline gelmiştir. Bunun nedeni kısmen liberal varsayımları, genellenebilirliği ve hesaplama kolaylığıdır. Katsayı a, test formlarının (hatta tek tek maddelerin) esasen tau-eşdeğer olduğunu varsayar; bu da aynı ölçekte aynı yapıyı ölçtükleri, ancak ortalamalarının ve standart sapmalarının değişebileceği anlamına gelir (Graham, 2006, s.930). Normatif temelli ölçekleme kullanılarak oluşturulan ölçek puanlarının güvenilirliğini tahmin etmek için Cronbach Alpha katsayısı kullanılabilir. Normatif temelli ölçek puanlarında araştırmacının ilgisi, sınava giren kişinin ilişkili olduğu norm grubunun performansına göre sınava giren kişilerin göreceli sıralamasına odaklanır (Almehrizi, 2013, s.443)

1.3.Geniş ölçekli sınavlar

Uluslararası büyük ölçekli değerlendirmeler (ILSA), ülkeler arasında eğitim çıktılarının genelleştirilmiş tanımlarını yapmak için kullanılabilecek

veriler üretir. Ayrıca eğitim sistemlerinin farklı aşamalarında ve seviyelerinde içerik bilgisi toplarlar (Strietholt ve Scherer, 2016, s.1).

Uluslararası geniş ölçekli sınavlar (ILSA) açıkça politikayı etkilemeyi amaçlamaktadır. Bunu yapmanın bir yolu, doğrudan politika yapıcıları ve hükümet yetkililerini hedef alan raporlar ve sunumlardır. Bir diğer yol ise bir dizi medya kuruluşu aracılığıyla kamusal alanda bilgi yaymaktır (Hamilton, 2017, s.281). Uluslararası geniş ölçekli sınavların temel özelliği, ülkeler arasında karşılaştırılabilirlik sağlamak üzere tasarlanmış olmalarıdır. Karşılaştırılabilirlik, incelenen nüfusa ve araçların kültürler arası geçerliliğine sahip olma durumudur. İlgili örneklemler hem yaş hem de okullaşma açısından dengeli değilse uluslararası karşılaştırmalar yanlılık içerir (Strietholt, Rosén, ve Bos, 2013, s.2). Ekonomik İşbirliği ve Kalkınma Örgütü'nün (OECD) etkili Uluslararası Öğrenci Değerlendirme Programı (PISA) gibi okul sektöründeki uluslararası karşılaştırmalar iyi bilinmekle birlikte, bu tür çalışmaların her biri kendi özellikleri ve metodolojileri ile erken yaşta eğitim, mesleki beceri eğitimi, yetişkin hayat boyu öğrenme ve yükseköğretim dahil olmak üzere diğer eğitim sektörlerini etkilemeyi amaçlamaktadır (Hamilton, 2017, s.284).

2.PISA

PISA, eğitim sistemlerinin kalitesini öğrenci sonuçları açısından izleyen uluslararası standartlaştırılmış bir değerlendirmedir. PISA, öğrencilerin topluma katılımları için gerekli bilgi ve becerileri ne kadar iyi edindikleri hakkında bilgi toplamak amacıyla 15 yaşındaki öğrencilerin becerilerini değerlendirmektedir (Davier, Gonzalez, Kirsch ve Yamamato, 2013, s.4). PISA, öğrencilerin okuma, matematik, fen bilimleri ve problem çözme alanlarındaki bilgi ve okuryazarlık becerilerini ölçmektedir. PISA yalnızca bu dört akademik alandaki öğrenci ustalığını ölçmekle kalmayıp, öğrencilerin gelecekteki zorluklarla başa çıkmak için kullandığı bilgi ve becerileri kullanma düzeylerini de belirler (Simon, Ercikan ve Rousseau, 2013, s.14). PISA araştırmaları her üç yılda bir yapılmaktadır. Her değerlendirme için okuma, matematik veya fen ana alan olarak seçilmekte ve kalan iki yan alana göre daha fazla önem verilmektedir. 2000, 2009 ve 2018 yıllarında ana alan okuma; 2003 ve 2012 yıllarında matematik; 2006 ve 2015 yıllarında ise fen olmuştur (OECD, 2019a, s.2).

PISA, öğrenci performansını değerlendiren ve performans farklılıklarını açıklamaya yardımcı olabilecek öğrenci, aile ve kurumsal

faktörler hakkında veri toplayan en kapsamlı ve titiz uluslararası programdır. Değerlendirmelerin kapsamı ve niteliği ile toplanacak arka plan bilgilerine ilişkin kararlar, katılımcı ülkelerin önde gelen uzmanları tarafından alınır ve hükümetler tarafından ortak, politika odaklı çıkarlar temelinde birlikte yönlendirilir. Değerlendirme materyallerinde kültürel ve dilsel genişlik ve dengeyi sağlamak için önemli çaba ve kaynaklar ayrılmıştır. Çeviri, örnekleme ve veri toplama süreçlerinde sıkı kalite güvence mekanizmaları uygulanmaktadır. Sonuç olarak, PISA sonuçları yüksek derecede geçerlilik ve güvenilirliğe sahiptir (OECD, 2019b, s.13)

2.1.PISA ve Geçerlik

Öğrenci geçmişi, uygulamaları, tutumları ve algılarına ilişkin karşılaştırılabilir ölçümlerin geliştirilmesi PISA'nın başlıca hedeflerinden biridir. Anketlerden elde edilen ölçümler genellikle ülke içinde ve ülkeler arasında öğrenci performansındaki farklılıkları tahmin etmek için kullanıldığından ve bu nedenle eğitim sistemlerini iyileştirme yolları hakkında politika ile ilgili potansiyel bilgi kaynakları olduğundan, bu yapıların ülkeler arası geçerliliği özellikle önemlidir (Avvisati, Le Donné ve Paccagnella, 2019, s.7). Geçerliğin kontrolü için yapılan çalışmalardan birisi iç tutarlılık yöntemidir. Her bir ölçeğin ülkeler içindeki iç tutarlılığını kontrol etmek ve ülkeler arasında karşılaştırmak için Cronbach Alpha katsayısı kullanılmıştır. Katsayı 0 ile 1 arasında değişmekte olup, yüksek değerler daha yüksek iç tutarlılığa işaret etmekte ve madde setinin ortak bir boyutu yakından ölçtüğü anlamına gelmektedir. Yaygın olarak kabul edilen kesme değerleri mükemmel için 0.9, iyi için 0.8 ve kabul edilebilir iç tutarlılık için 0.7'dir. Bazı ölçekler için bazı ülkeler bir veya iki maddeyi silinmiştir (OECD, 2019a, s.8).

2.2.PISA ve Güvenirlik

PISA uygulamasında farklı güvenirlik sağlama yöntemleri kullanılmaktadır. Bunlardan biri de kodlayıcı güvenirliği oalrak söylenebilir. Kodlayıcı güvenirliği, bir ülke içindeki değerlendirme sonuçlarının geçerliliğinin yanı sıra ülkeler arasında değerlendirme sonuçlarının karşılaştırılabilirliğini sağlamak için kritik öneme sahiptir. Kodlayıcı güvenilirliğinin değerlendirilmesi çoklu kodlama tasarımı ile mümkün olmuştur. Ülke içi kodlayıcı güvenilirliğini değerlendirmenin amacı, bir ülke içinde kodlama güvenilirliğini sağlamak ve puanlama sürecindeki herhangi bir kodlama tutarsızlığını veya sorunu tespit etmek ve böylece sürecin daha erken aşamalarında ele alınıp çözülebilmelerini sağlamaktır. PISA gibi uluslararası

büyük ölçekli bir değerlendirmede tüm öğrenci yanıtlarını çoklu kodlamak ekonomik değildir, bu nedenle ulusal maliyetleri ve kodlayıcı yükünü azaltmak için çoklu kodlama ve tekli kodlamayı birleştiren bir kodlama tasarımı kullanılmıştır (Organisation for Economic Co-Operation and Development, 2016, s.8). Bir yanıt tekli kodlandığında, kodlayıcılar doğrudan kitapçıklarda işaretleme yapar. Bir yanıt çoklu kodlandığında, son kodlayıcı doğrudan kitapçıkta kodlama yaparken diğerleri kodlama sayfalarında kodlama yapar; bu, kodlayıcıların kodlama kararlarında bağımsız kalmalarını ve kodlama güvenilirliğinin doğru bir şekilde değerlendirilmesini sağlar (OECD, 2019a, s.7).

3.TARTIŞMA VE SONUÇ

PISA uygulamalarının geçerlik ve güvenirliğini belirlemek amacıyla alanyazında farklı çalışmalar yer almaktadır. Söz konusu çalışmalardan birisi Kılıç, Erdem-Kara ve Doğan (2019) tarafından yapılan PISA 2012 uygulamasının ölçüt geçerliğini beklenti tabloları ve uyum analizi kullanarak yaptıkları araştırmadır. Elde edilen bulgulara göre PISA 2012'de alt ve üst kategorilerdeki uyumun 2003'e göre daha yüksek olduğu, her iki yıl için de genel anlamda matematik okuryazarlığının, problem çözme becerileriyle benzer olduğu, diğer bir deyişle matematik okuryazarlığının zamandaş geçerliğinin sağlandığı sonucuna ulaşmışlardır.

Polat, Toraman ve Sölpük-Turhan (2022) tarafından yapılan başka bir çalışmada ise PISA 2018 Türkiye verileri üzerinde yapılan madde analizlerine göre Okuma Yazma Öğrenci Anketi'nin (OOYÖA) geçerlik ve güvenirlik düzeylerini MTK temelinde ortaya koymak amaçlanmıştır. Bu amaçla, PISA 2018 öğrenci anketinde yer alan üç soru grubu birbirinden bağımsız olarak yapılandırılmış ve her bir soru grubu bağımsız bir ölçek olarak kabul edilmiştir. ST164, ST165 ve ST166 soru setleri MTK'ya göre analiz edilerek ST164, ST165 ve ST166 soru setleri üzerinde yapılan madde korelasyon matrisi analizi sonucunda, tek boyutlu bir yapının olmadığı ve Q3 testine göre anketteki maddeler arasında yerel bağımsızlığı bozan hiçbir maddenin bulunmadığı sonucuna varılmıştır.

Aditomo ve Köhler (2020) tarafından yapılan başka bir çalışmada ise öğrenci derecelendirmelerinin okul düzeyinde toplanması nedeniyle PISA 2015 tarafından sağlanan sınıf yönetimi, duygusal destek, sorgulamaya dayalı öğretim, öğretmen odaklı öğretim, uyarlanabilir öğretim ve geri bildirimi ölçen altı ölçek için geçerli ve güvenilir öğretim kalitesi sağlayıp

sağlamadığını araştırmışlardır. Elde edilen bulgulara göre çoğu ülkede/bölgede, okul düzeyindeki güvenirlik sınıf yönetimi ölçeği için yeterli bulunurken, diğer ölçekler için düşük ya da orta düzeyde bulunmuştur. Faktöriyel ve yordayıcı geçerlilik incelendiğinde, sınıf yönetimi, duygusal destek, uyarlanabilir öğretim ve öğretmen yönlendirmeli öğretim ölçeklerinin okullar arasındaki öğretim kalitesinde anlamlı farklılıklar yakaladığı görülmüştür. Bu arada, sorgulama ölçeği neredeyse tüm ülkelerde/bölgelerde zayıf geçerlilik sergilemiştir.

Rutkowski ve Rutkowski (2013) tarafından PISA 2009'da sosyoekonomik göstergenin önemli bir bileşeni olan ev eşyaları endeksini ölçüldüğü veriler üzerinde güvenilirliğini ve geçerliliğinin bazı yönlerini araştırmışlardır. Araştırma bulgularına göre, mevcut endeksle ilgili kayda değer endişeler olduğunu göstermektedir; bunlar arasında ülkelere göre oldukça değişken güvenilirlik, bazı alt ölçeklerde model-veri tutarlılığının zayıf olması ve kültürel karşılaştırılabilirliğin zayıf olduğuna dair kanıtlar yer almaktadır.

Bu araştırma PISA sınavının farklı yıllara göre geçerlik ve güvenirliğini inceleme amacıyla yapılmıştır. Elde edilen bulgulara göre PISA uygulamasının geçerliği ve güvenirliği ülkelere göre farlılık göstermektedir. Ayrıca PISA uygulamasının alt boyutları da geçerlik ve güvenirlik anlamında değişkenlik olduğu görülmektedir. Bundan sonraki çalışmalar için ölçme değişmezliği gibi bir konuda farklı ülkeler arasında karşılaştırmalar yapılabilir. Ayrıca matematik, fen bilimleri ve okuma alanlarında her boyut ele alınarak geçerli ve güvenilir olma durumu saptanabilir.

KAYNAKÇA

Aditomo, A. ve Köhler, C . (2020). Do student ratings provide reliable and valid information about teaching quality at the school level? Evaluating measures of science teaching in PISA 2015, *Educational Assessment, Evaluation and Accountability*. https://doi.org/10.1007/s11092-020-09328-6

Almehrizi, R. S. (2013). Coefficient Alpha and Reliability of Scale Scores, *Applied Psychological Measurement, 37*:438. https://doi.org/10.1177/0146621613484983

Avvisati, F., Le Donné, N. ve Paccagnella, M. (2019). A meeting report: cross-cultural comparability of questionnaire measures in large-scale international surveys. *Measurement Instruments for the Social Sciences*, 1-8. https://doi.org/10.1186/s42409-019-0010-z

Baykul, Y. ve Turgut, M. F. (2010). Eğitimde Ölçme ve Değerlendirme. Ankara: Pegem Akademi Yayıncılık.

Bollen, K. A. (1989). *Structural Equations with Latent Variables*, New York: Wiley.

Brennan, R. L. (2011). Generalizability Theory and Classical Test Theory, *Applied Measurement in Education, 24*, 1–21. https://doi.org/10.1080/08957347.2011.532417

Büyüköztürk, Ş. (2009). *Sosyal bilimlerde için veri analizi el kitabı*, Ankara: Pegem Akadmei Yayıncılık.

Crocker, L. ve Algina, J. (2008). *Introduction to Classical and Modern Test Theory,* Cengage Learning.

Davier, M., Gonzalez, E., Kirsch, I ve Yamamoto, K. (2013). *The Role of International Large-Scale Assessments: Perspectives from Technology, Economy, and Educational Research*, New York: Springer.

Drost, E. (2011). Validity and Reliability in Social Science Research, *Education Research and Perspectives, 38,* 1.

Crocker, L. ve Algina, J. (2008). *Introduction to Classical and Modern Test Theory,* Cengage Learning.

Ebel, R. L. (1961). Must all tests be valid? *American Psychologist, 16*, 640-647.

Ercan, İ. ve Kan, İ. (2004). Ölçeklerde Güvenirlik ve Geçerlik, *Uludağ Üniversitesi Tıp Fakültesi Dergisi, 30 (*3) 211-216.

Graham, J. M. (2006). Congeneric and (Essentially) Tau-Equivalent Estimates of Score Reliability What They Are and How to Use Them, *Educational and Psychological Measurement, 66*(6), 930-944. https://doi.org/10.1177/0013164406288165

Hamilton, M. (2017). How International Large-Scale Skills Assessments engage with national actors: mobilising networks through policy, media and public knowledge, *Critical Studies in Education*, 58, 3, 280–294. https://doi.org/10.1080/17508487.2017.1330761

Haynes, S. N., Richard, D. C. S. ve Kubany, E. S. (1995). Content Validity in Psychological Assessment: A Functional Approach to Concepts and Methods, *Psychological Assessment,7(*3), 238-247.

Kalaycı, Ş. (2014). *SPSS Uygulamalı çok değişkenli istatistik teknikleri*, Ankara: Asil Yayıncılık.

Kılıç, A. F., Erdem-Kara, B. ve Doğan, N. (2019). PISA 2003 ve 2012 Matematik Okuryazarlığı Puanlarının Ölçüt Geçerliği: Beklenti Tabloları ve Uyum Analizi, *Eskişehir Osmangazi Üniversitesi Sosyal Bilimler Dergisi, 20*. https://doi.org/10.17494/ogusbd.

Messick, S.(1989).*Validity.* In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 13-103). New York: Macmillan.

Messick, S. (1994). Validity of Psychological Assessment: Validation of Inferences From Persons' Responses and Performances as Scientific Inquiry into Score Meaning, *Educational Testing Service,*Princeton N. J.

Nunnally, J. C. ve Bernstein, I. H. (1994). *Psychometric Theory*, New York: Mc Graw-Hill.

Organisation for Economic Co-Operation and Development. (2016). *PISA 2015 Technical Report,* Paris: OECD Publishing.

Organisation for Economic Co-Operation and Development. (2019a). *PISA 2018 Technical Report,* Paris: OECD Publishing.

Organisation for Economic Co-Operation and Development. (2019b). *PISA 2018 Assessment And Analytıcal Framework,* Paris: OECD Publishing.

Polat, M., Toraman, Ç. ve Sölpük-Turhan, N. (2022). Reliability analysis of PISA 2018 reading literacy student questionnaire based on Item Response Theory (IRT): Turkey sample, *International Journal of Curriculum and Instruction, 14(*1), 1004–1028.

Rutkowski, D. ve Rutkowski, L. (2013). Measuring Socioeconomic Background in PISA: one size might not fit all, *Research in*

*Comparative and International Education, 8(*3). http://dx.doi.org/10.2304/rcie.2013.8.3.259

Scholtes, V. A., Terwee, C. B. ve Poolman, R. W. (2011). What makes a measurement instrument valid and reliable? *Injury, Int. J. Care Injured, 42,* 236–240. https://doi.org/10.1016/j.injury.2010.11.042

Simon, M., Ercikan, M. K. ve Rousseau, M. (2013). *Improving Large-Scale Assessment in Education Theory, Issues, and Practice,* London: Routledge.

Singh, A.S. ve Masuku, M. B. (2012). Understanding and applications of test characteristics and basic inferential statistics in hypothesis testing, *European Jr. of Applied Sciences, 4*(2), 90-97.

Strietholt, R. ve Scherer, R. (2016). The Contribution of International Large-Scale Assessments to Educational Research: Combining Individual and Institutional Data Sources, *Scandınavıan Journal Of Educatıonal Research,* http://dx.doi.org/10.1080/00313831.2016.1258729

Strietholt, R., Rosén, M., & Bos, W. (2013). A correction model for differences in the sample compositions: The degree of comparability as a function of age and schooling? *Large-Scale Assessments in Education, 1*(1), 1–20. http://dx.doi.org/10.1186/2196- 0739-1-1

Tekin, H. (1991). *Eğitimde Ölçme ve Değerlendirme,* Ankara: Yargı Yayıncılık.

Wagemaker, H. (2020). *Reliability and Validity of International Large-Scale Assessment Understanding IEA's Comparative Studies of Student Achievement,* New Zealand: Springer https://doi.org/10.1007/978-3-030-53081-5