Değerlendirme ve Geçerlik Üzerine Arkeolojik Bir Çaba – Bağlam ve Kavram Bilgisi

An Archaeological Effort on Assessment and Validity - Contextual and Conceptual Knowledge

Sıla Elif Törün

Ege Üniversitesi Tıp Fakültesi Tıp Eğitimi Anabilim Dalı

Anahtar Sözcükler: Değerlendirme, ölçme, geçerlik, pozitivizm, eleştirel refleksiyon

Key Words: *Assessment, measurement, validity, positivism, critical reflection*

ÖZET: Öğrenmenin ve öğrencilerin değerlendirilmesine ilişkin her uygulama bilimsel ve felsefi temel varsayımlara dayanmaktadır. Değerlendirme konusunda sistematik bir sorgulama yapılabilmesi, bu konuda yaşanan sorunlara çözüm aranması, yöntem ve teknik tartışmalarının ötesinde, bu temel varsayımlarla değerlendirme uygulamaları arasındaki ilişkinin dikkate alınmasını gerektirir. Tarihsel süreç göz önünde bulundurulduğunda, değerlendirme kavramsallaştırması ve söyleminin, sosyal, politik, ekonomik ve kültürel bağlama dayalı olarak ortaya çıkan bu temel varsayımlar ve ihtiyaçlar üzerinde geliştiği görülmektedir. Geçerlik ve geçerliğin gösterilmesine ilişkin tartışmaların da değerlendirme konusu ile beraber bütünlük içinde ele alınması, geçerlik kavramına ilişkin daha derin bir kavrayışın gelişmesine yardımcı olacaktır.

Değerlendirme ve geçerlik uygulamalarına ilişkin günümüzde kullanılan bilimsel söylemi ve ölçme ağırlıklı değerlendirme anlayışını, ortaya çıktığı tarihsel bağlam ve temel varsayımlar çerçevesinde ele almayı amaçlayan bu makale, Foucault'nun öne sürdüğü ve "arkeoloji" olarak adlandırdığı tarihsel analiz yönteminden esinlenerek hazırlanmıştır. Değerlendirme ve geçerlik konularına ilişkin bilimsel söylemi bu çerçevede sorgulayarak kendi kavramsallaştırmamızı, değerlendirme ve geçerlik uygulamalarını yeni sorularla başka bir açıdan gözden geçirmek, ağırlıklı olarak yöntem ve teknikler üzerine yürütülmekte olan tartışmalara da farklı bir boyut kazandırabilir.

ABSTRACT: *Praxis of educational assessment is based on a set of scientific and philosophical assumptions. Beyond the discussions related to technical concerns, a systematic inquiry on assessment, and seeking of solutions for problems on assessment issues require one to take the relationship between these assumptions and assessment practices into consideration. Considering the historical background, it becomes explicit that conceptualization of assessment and discourse on assessment have developed based on these assumptions and needs which had been emerged in consequence of the social, political, economical, and cultural context. Approaching to the validity and validation issues with such a comprehensive perspective in terms of assessment will also lead up to a deeper comprehension on the concept of validity. This paper, aiming to address contemporary scientific discourse and measurement based assessment approach within the framework of the basic assumptions and historical context in which those had arise, is inspired by Foucault's historical analysis method referred as "archeology". This paper also aims to offer an in-depth examination of today's dominant*

scientific discourse on assessment and validity through questioning our own conceptualization and practices; thus adding another dimension to the discussions mostly focusing on methods and technical issues.

Yazışma Adresi: Ege Üniversitesi Tıp Fakültesi Tıp Eğitimi Anabilim Dalı, 35100, Bornova-İzmir
e-posta: silaelif.torun@ege.edu.tr

Giriş

Öğrenme ve öğrencilere ilişkin değerlendirme, eğitim alanında her dönemde ve her öğretim kademesinde en çok tartışılan konulardan birisi olmuştur. İlgili yayınlarda, resmi belgelerde, yürütülen uygulama ve tartışmalarda ülkemizde genellikle "ölçme-değerlendirme" başlığı altında ele alınan ve bu konuyla ilişkili kaynaklarda tanımı, "bir ölçme sonucunu bir ölçütle karşılaştırarak bir yargıya varmak" olarak yapılan (1-4) değerlendirme, bu adlandırma ve tanımlamaya göre, doğrudan ölçme sonuçlarına dayandırılmıştır. Stevens tarafından 1951 yılında, "nesnelere ve olaylara belli bir kurala göre sayısal nitelikte değerlerin verilmesi" olarak yapılan tanım, ölçmeye ilişkin en çok kullanılan tanımlardan birisidir (5,6). Eğitim bağlamında "assessment" karşılığı olarak dilimizde "ölçme-değerlendirme" başlığının kullanılması, yapılan işi tanımlamak ve program değerlendirme ya da genel olarak değerlendirme ile öğrenme ve öğrenciye ilişkin değerlendirmeleri ayırmak için anlam birliği ve kolaylık sağlamaktadır. Bununla beraber, "ölçme-değerlendirme" adlandırması, değerlendirmeye ilişkin temel bir "söylem" olarak, bir kavramsallaştırmanın da ifadesidir. Psikometrik açıdan kuşkusuz doğru olan bu kullanım ve tanımlama, öğrenme ve öğrenciye ilişkin değerlendirme kavramsallaştırması ve temel varsayımları açısından farklı biçimde ele alınıp tartışılabilir.

Linn ve Miller (7), eğitim alanında değerlendirmeyi (assessment), "öğrenmenin yapı ve kapsamını belirleme amacıyla öğrenme ve öğrenilenlere ilişkin bilgi edinmek için çeşitli işlemlerin kullanıldığı bir süreç" olarak tanımlamışlardır. Benzer tanımlar üzerinden, çok sayıda yazar, değerlendirmeyi ölçmenin dışında ve ölçmeden daha geniş bir kavram olarak ele almış; ölçmenin, değerlendirme yapabilmenin yollarından birisi olduğunu vurgulamışlardır (7-13).

Değerlendirme ve ölçme ilişkisini tartışırken, öğrenmenin ve öğrencilerin değerlendirilmesine ilişkin her uygulamanın ve bu uygulamalara ait geçerlik çalışmalarının bazı varsayımlara dayandığını dikkate almak gerekir. Pellegrino, Chudowsky ve Glaser (14), değerlendirmenin dayandığı bilimsel ve felsefi temel varsayımları üç başlıkta ele almıştır. Bu varsayımlardan ilki öğrenmeye ilişkindir; değerlendirme yaklaşımı ve uygulamalarımız, bilginin ne olduğu ve öğrenmenin nasıl gerçekleştiğine dair açıklamalarımız çerçevesinde biçimlenir. İkincisi, öğrenmenin ve öğrencilerin değerlendirilmesi için önemli ve değerli olan bilgilerin ne tür gözlemler ya da hangi yollarla elde edilebileceğine ilişkin varsayımlarımızdır. Üçüncü temel varsayım ise, değerlendirme kapsamında, öğrenme ve öğrencilere ilişkin edinilmiş bilgileri, anlamlı çıkarımlar yapmak için en iyi biçimde nasıl yorumlayıp kullanacağımız ile ilgilidir.

Değerlendirmenin "geçerlik" konusuyla beraber ele alınması kaçınılmazdır ve değerlendirme ile ilgili tüm varsayımlar geçerlik tartışmalarının da merkezindedir. Kavramsal olarak geçerliğin tanımlanması (validity) ile uygulama açısından geçerliğin gösterilmesi (validation) arasında doğrudan bir ilişki vardır; geçerliğin nasıl gösterileceği, geçerliğin nasıl tanımlandığına bağlıdır. Geçerliğe ilişkin yapılan tanımlama, bu kavramın ne için kullanıldığı, ne anlama geldiği, sınırlarının ne olduğu ve diğer kavramlarla ilişkisinin nasıl olduğu sorularına cevap verir; geçerliğin gösterilmesi ise, ne tür ve ne ölçüdeki kanıtın hangi gerekçeyle, nasıl analiz edileceği gibi uygulama alanında soracağımız soruları yanıtlamamızı gerektirir (15).

Tarihsel süreç ve bağlam göz önünde bulundurularak incelendiğinde, değerlendirme yaklaşım ve uygulamalarına ilişkin gelişmelerle geçerlik kavramına dair tartışmalar arasındaki yakın ilişki dikkat çekicidir; geçerlik konusunun temel kodları, değerlendirme yaklaşım ve uygulamalarının tarihsel süreçteki seyri içinde yer almaktadır. Değerlendirme ve geçerlik konularının, birbirini karşılıklı olarak etkileyen, dönüştürüp değiştiren bu ilişki bağlamında bir arada ele alınması, geçerlikle ilgili tanım ve işlemlerin kuramsal zeminini, bağlamını ve bu konuda yürütülen tartışmaları daha iyi kavramaya da katkıda bulunabilir.

Değerlendirme ve geçerliğe ilişkin varsayımların ve bunların üzerinde şekillenen söylem ve uygulamaların, sosyal, kültürel, ekonomik ve politik gelişmelerle biçimlendiği ve ortaya çıkan farklı ihtiyaçlara cevap verecek şekilde değiştiği görülmektedir (9,11,16,17). Bu noktada, değerlendirme ve geçerlik konularıyla ilgili "bugünü" anlamak ve anlamlandırmak için geçmişin bilinmesinin önemi ortaya çıkmaktadır. Düz bir tarihçeye dayalı geçmiş bilgisi ile sınırlı olmayan "tarih bilgisi", kavram, kural ve uygulamaların ortaya çıkışını belirleyen koşulları, sosyoloji, ekonomi, psikoloji, eğitim gibi farklı alanlar arasındaki ilişkileri kavramaya yönelik bir çabayı içerir.

Değerlendirme ve geçerlik konularını bu çerçevede ele almaya çalışan bu yazı, Foucault'nun yaklaşımından esinlenerek hazırlanmıştır.

Yöntem

Geçtiğimiz yüzyılın en önemli düşünür ve tarihçileri arasında sayılan Michel Foucault (1926–1984), sadece tıp eğitiminde sık tartışılan profesyonalizm, yetkinlikler, standartlar, iletişim, değerlendirme gibi birçok konuya farklı bir anlam ve derinlik kazandırabilecek kavramsal katkısı ile değil, kullandığı yöntem açısından da tıp eğitimi alanında dikkate alınması gereken bir isimdir.

"Şimdinin tarihini" yazdığını ve görevinin "bugün açısından dünün tarihine bakmak" olduğunu belirten Foucault, bilgi ve düşünce sistemlerinin farklı dönemlerdeki izlerini sürerek günümüzde kullanılan bilimsel söylemin tarihsel analizini yapmıştır. Foucault, belli bir duruma ilişkin gerçeklik algımızın, o alana ilişkin bilimsel söylemler tarafından şekillendiğine ve bu söylemin düşünce ve eylemlerimizi etkileyip belirlediğine işaret etmiştir (18). Foucault'nun "arkeoloji" adını verdiği yöntem, "bir çağın söylemlerinin normatifliği, bilginin normalleşme biçimleri ve oluşum kurallarını inceleyen bir yöntemdir" ve "kavramların yer değiştirmesini ve dönüşmesini" anlamamıza yardımcı olur (18). Foucault'ya göre "arkeoloji" ile ulaşılan bilgi (savoir), bir uygulamanın, bir görüşün, bir kuramın belli bir anda ortaya çıkışını mümkün kılan koşullar, dönemin düşünce biçimi, felsefesi, siyasi ve ekonomik uygulamaları, toplumsal kurumların işleyişi, kısacası bilimsel söylemi belirleyen ve şekillendiren duruma ilişkin bilgidir (18,19). Bilimsel kitaplarda, kuramlarda yer almış, belli bir disipline ait formal ve bilimsel niteliğe sahip bilgiye dayanan söylemi "connaissance" sözcüğü ile ifade eden Foucault, bilimsel söylemin "savoir" içinde ortaya çıktığına işaret etmiştir; Foucault'a göre, daha geniş çerçevedeki "savoir" anlaşılmadan "connaissance"ın anlaşılabilmesi mümkün değildir (18,20). Foucault'nun "genealogy" (soy ağacı) adını verdiği diğer yöntem ise, bilgi ve bilimsel söylemin iktidar ve güç ilişkileri tarafından nasıl belirlendiğini, "doğrular" ve "yanlışların", her boyuttaki iktidar/güç mekanizmaları aracılığıyla farklı dönemlerde nasıl değiştiğini inceleyen, bilimsel bilgi ve söyleme dair eleştirel bir tarih analizi yapmanın yöntemidir (19,20).

Bu yazı, Foucault'nun kullandığı biçimiyle "arkeoloji" olarak nitelendirilebilecek kapsam ve derinlikten uzak olmakla birlikte, esin kaynağını Foucault'dan alan, değerlendirme ve geçerlik ile ilişkili bilimsel söylemin ortaya

çıktığı "savoir"ı anlama çabası olarak kabul edilebilir. Uygunluk ve doğruluğunu öncelikle kabul edip kendi alanımıza transfer ettiğimiz nesnellik, kriter, standart, ayırıcılık, ölçme-değerlendirme gibi çok sayıda kavram, işlem ve başlığı içeren değerlendirme alanındaki ölçüm odaklı "bilimsel söylemi", teknik-yöntem ve tanımlarla sınırlı kalmadan, tıp eğitimi bağlamı dışından farklı zaman ve alanlar üzerinden anlamak, günümüzdeki kavramsallaştırma ve uygulamalara değişik açılardan bakabilmemize katkıda bulunabilir.

Bilimsel söylemlere, kavram ve uygulamalara eleştirel bakış ve sorgulama, eleştirel refleksiyon gibi Tıp Eğitimi disiplini için özellikle önemli ve gerekli olan diğer adım ise, belki de önce her okurun "genealogy" ilhamıyla kendi tarih okumasını yapması ile atılabilir.

1. Değerlendirmenin Tarihsel Süreçte Değişen Anlam ve İşlevi

1.1. Ortaçağ (Kilise, aristokrasi, feodalizm)

Ortaçağ Batı dünyasında ağırlıklı olarak din etkisi altında olan "düşünce", dünyayı metafizik bir anlayışla açıklamıştır. Din ve dinin temsilcisi konumuna yerleştirilen Kilise, aristokrasi ile beraber, bireysel ve toplumsal ilişkiler için kural koyucu role sahip olmanın ötesinde, gerçeğin (realite) ve gerçeğe ulaşan yolun (bilgi) ne olduğunu belirleyen bir kurum niteliğiyle, bilim, eğitim gibi alanları da düzenlemiştir.

Günümüzdeki anlamıyla okullaşma, Batı'da 17. yüzyılda başlamıştır (21). Ortaçağda eğitim, feodal sistem bağlamında temel olarak din ve aristokrasi tarafından yürütülmüş, din adamlarının yetişmesi ve seçkinlerin bireysel niteliklerini geliştirmesi amacıyla işlev görmüştür (21,22). Ortaçağda Avrupa üniversitelerinde eğitime erişim daha çok ailevi ve sosyal statü tarafından belirlenmiştir (10,22,23). Bu dönemde eğitimde değerlendirme, ayrı bir uygulama olarak değil, öğretme görevi ve öğrenmenin içinde didaktik amaçla ve eğitsel işlevleriyle yer almıştır (22,23). Günlük öğrenme ve öğretme pratiği sırasında sözlü biçimde karşılıklı sorulan sorular, verilen cevaplar ve performansın eş zamanlı gözlemiyle öğrenme durumuna ilişkin niteliksel bir değerlendirme yapılmış; eğitim sırasında yazılı sınav ve notlandırma bu dönemde henüz kullanılmaya başlanmamıştır (22-24).

1.2. 18. yüzyıl (Aydınlanma çağı, modernite, merkantilizim)

Batı'da Rönesans ile başlayıp 18. yüzyıl boyunca süren ve Aydınlanma çağı olarak adlandırılan dönemde modernizmin ilk temelleri atılmış, akıl ve bilim ilerlemenin araçları olarak görülmüştür. Önceki yüzyıllarda temel olarak din ve aristokrasi eliyle yürütülen eğitim, 18. yüzyılda Batı'da demokrasi ve eşitlik kavramları çerçevesinde kurumsallaşmaya başlamış; okuma, yazma ve basit aritmetik işlemlerini kapsayan temel eğitim geniş toplum kesimlerine yayılmıştır (21). Eğitimin çocukluk çağında başlaması, yaş ve süreye göre sınıfların oluşturulması, "sınıf geçme"nin tanımlanması bu yüzyılda başlayan uygulamalar olmuştur (22).

Önceki yüzyıllarda keşifler ve sömürgeciliğin de etkisiyle ticaretin gelişmesi, birikimlerin ticari ve mali sermayeye dönüşmesine yol açmış ve geniş toprak sahipliğine dayalı feodalizm, 18. yüzyıla gelindiğinde Batı dünyasında yerini ticari kapitalizme bırakmıştır. Ticarete dayalı ekonomi üzerinde gelişen ticaret burjuvazisi ile birlikte hukuk, bankacılık, ticaret, taşımacılık gibi iş alanları ortaya çıkmıştır. Bu gelişmelerle beraber iş yaşamında önemli pozisyonlara ve eğitime erişimde aile ve sosyal statüden kaynaklanan ayrıcalıklara karşı fırsat eşitliği ve liyakat savunulmuştur (10,22,24,25). Ancak 18. yüzyıl boyunca, geniş yoksul nüfusun temel okuryazarlık eğitimi dışında üniversite düzeyindeki akademik eğitime erişimi sınırlı kalmış, bu düzeydeki eğitimin maliyetini karşılayabilmek, ağırlıklı olarak dönemin seçkinleri ve yüksek gelir düzeyine sahip tüccar

aileleri için mümkün olmuştur (21,22).

Üniversite düzeyindeki eğitimde öğrenme sürecinin ve öğrenilenlerin niteliğinin değerlendirilmesinde temel yöntem olarak, sözlü sınavlar ve münazaralar (disputations) kullanılmıştır (22,26). Öğrencilerin, kendilerini yetiştiren eğiticiler ile birlikte, başka bir eğiticinin yönetiminde, belli bir konu hakkında kendilerine verilen bir tezi diğer bir grup karşısında savundukları bu münazaralar için sayısal bir not sistemi kullanılmamış; tartışma içeriği, savunulan argümanlar ve gösterilen performans niteliksel olarak değerlendirilmiştir (22-24,26). Eğiticiler, öğrencilerini bu tartışmaya hazırlamak ve tartışma sonrası savunmaya ilişkin güçlü ve zayıf noktaları, tartışma konusunu ya da problemin çözümünü özetleyip bildirmekle sorumlu tutulmuştur (22,24). Diğer öğrencilere ve halka açık yapılan bu münazaralar, hem tartışanlar hem de izleyenler için bir öğrenme fırsatı yaratmasının yanı sıra, yerleşik bilginin kullanılması, eleştirel analizi ve yeni bilgi üretimi amacıyla yapılmış; münazaralar aracılığı ile toplum gözünde üniversitelerin ve bu kurumlarda üretilen ve öğretilen bilginin meşruiyeti sağlanmıştır (22,23).

1.3. On dokuzuncu yüzyıl ve 20. yüzyıl ilk yarısı (Sanayileşme, pozitivizm, verimlilik)

On sekizinci yüzyıl sonu ve 19. yüzyıl, Avrupa ve Amerika'da modernite çerçevesinde ulus devletlerin tarihte yerini almaya başladığı dönem olmuştur. Bu dönemde vatandaşlık temelinde ulus inşa etmenin en önemli araçlarından birisi olarak eğitim, ortak değer, anlayış ve entelektüel birikime sahip vatandaşların yetiştirilmesi amacıyla devlet örgütlenmesi içinde okullar aracılığıyla sistematik olarak kurumsallaşmıştır (21). Sanayi Devrimi ile beraber merkantalizmden sanayi kapitalizmine geçen Batı Avrupa ve ABD'de, önceden çoğunluğu tarım alanında çalışan ve dağınık kırsal yerleşime sahip nüfus, 19. yüzyılda kurulan çok sayıdaki fabrika ve işletmelerde çalışmaya başlamış ve nüfusun belli bölgelerde yoğunlaşması sonucu artan kentleşme ile birlikte öğretim kitleselleşmiştir. Yaş ve süreye göre belirlenen sınıflardan oluşan okullar ve yasal düzenlemelerle eğitim, bürokratik ve hukuki nitelik kazanmıştır (21,27). On dokuzuncu yüzyılın ilk yarısı boyunca değerlendirme, önceden olduğu gibi, öğrenme ve öğretme pratiğinin içinde soru-cevap biçiminde yer almış, öğrenme ve öğretmeden ayrı bir kavram olmayıp didaktik ve eğitsel işlevleriyle kullanılmıştır (22).

On dokuzuncu yüzyılın ikinci yarısında, kapitalizmin yükselmesi, sanayi burjuvazisi ve orta sınıfın gelişmesiyle, ekonomik sistem ve sosyal yapı da değişmeye başlamıştır. Bu dönemde, toplumsal iş bölümü içinde özel nitelik gerektiren işler ayrışmış, değişik profesyonel roller ortaya çıkmıştır (28,29). Schön'e göre, pozitivist epistemoloji çerçevesinde, mesleki uygulamalarda "sanat" ve "ustalığın" yerini bilimsellik almış; deneyler yoluyla elde edilmiş, güvenilir bilimsel bilgilerin kullanımına dayalı "teknik rasyonalite" (technical rationality) yaklaşımı bu dönemde ortaya çıkarak 20. yüzyıl ilk yarısı boyunca yaygınlaşmıştır (1). Bu anlayışın prototipi sayılan hekimlik ve mühendisliğin yanısıra, hukuk, veterinerlik, bankacılık gibi alanlar, özel kuramsal bilgi temelinde uygulanabilen, belli bir eğitim ve diplomayı gerektiren, alana özgü otonomi, yetki ve statü tanıyan özellikleriyle diğer işlerden ayrılmış ve modern anlamda "mesleklere" dönüşmüştür (27,29-31). Çalışma hayatında önemli pozisyonlara gelebilmede ailevi statü ve himayenin rolü azalarak bu meslekler aracılığıyla seçkin bir sosyal statü ve yüksek gelire sahip olmak mümkün hale gelmiştir (27,29). Sağladıkları imkânlar nedeniyle bu alanlardaki eğitime talep artmış (28,29); bu dönemde sınavlar, kişilerin gelecekleri ve kariyerleri için büyük önem kazanmıştır (22,28,29).

Eğitimin kitleselleşmesi, bireysel kazanımların önem kazanması ve başarılı olma isteği, eğitim alanı ve çalışma hayatında rekabetçi bir ortam yaratmış; tek tek yapılan değerlendirmeler yerine, çok sayıda kişinin aynı anda değerlendirilmesine imkân verecek "objektif" sınavlara duyulan ihtiyaçla birlikte, sözlü sınavlar yerini büyük ölçüde standardize, yazılı sınavlara bırakmaya başlamıştır (22,23,28,32).

Eğitim alanında yazılı sınavlar, ilk kez Oxford ve Cambridge Üniversitelerinde 18. yüzyıl sonunda kullanılmaya başlanmış olmakla beraber, belli bir dönem bu iki İngiliz Üniversitesi ile sınırlı kalan bu uygulamanın, Batı dünyasında yaygınlaşıp tüm eğitim kademelerinde kullanılmaya başlandığı dönem 19. yüzyıl olmuştur (10,22,26,31). ABD'de ilk yazılı sınav, 1845 yılında yapılmıştır (32-34). Madaus ve Stufflebeam'e göre (35), ABD'de sözlü sınavların yerini yazılı sınavlara bırakmasının başlangıcı olan bu uygulama, okulların ve eğitim programlarının değerlendirilmesinde sınav sonuçlarının temel kaynak olarak alındığı uzun sürecek bir geleneği de başlatmıştır. Değerlendirmenin sayısallaştırılması, hesaplanması (notlandırma), belgelendirilmesi ve saklanması, başka verilerle karşılaştırılması, sınıflama, kategorize etme, tek doğru cevabın belirlenmesi gibi uygulamalar, yazılı sınavlarla beraber ortaya çıkmıştır (22,24,26).

On dokuzuncu yüzyıl boyunca ortaya çıkan sosyal, politik ve ekonomik değişimler, eğitimin amacının, okul kültürünün ve değerlendirme uygulamalarının değişip dönüşmesine yol açmış (28), 20. yüzyılın ilk yarısında sanayileşme ve "verimlilik" (efficiency) düşüncesi, eğitim politikalarını ve uygulamalarını da doğrudan etkilemiştir (2). Verimlilik temelinde, işyerlerinde görev analizleri yapılıp standardizasyon öne çıkarken, okullarda da benzer biçimde öğrenme hedefleri tanımlanmış, eğitimde standardizasyon önem kazanmış ve hedefler doğrultusunda "ustalığın" kesin biçimde kazanılıp kazanılmadığını belirlemek, sınıf geçmek, diploma almak üzere sınavlar hazırlanmıştır (36). Verimliliğe dayalı eğitim programları, davranışçı öğrenme kuramları ve standardize, nesnel, "bilimsel yöntemlerle" yapılan "ölçümler" yoluyla öğrencilerin sıralanmasına dayalı değerlendirme anlayışı, okullardaki öğretimi biçimlendirmiştir (35,36). Bu dönemde okulların, değerlendirme sistemlerinin, seçme ve mezuniyet koşullarının ekonomik ve sosyal sermayenin dağıtılmasında düzenleyici rolü ortaya çıkmaya başlamıştır(22,23,37). Kişiler arası kapasite farklılıklarının ortaya çıkarılması ve bireylerin kapasitelerine göre sıralanması, seçim yapacak yöneticiler için önemli hale gelirken, kişiler için de "diğerlerine göre" kendi durumunu ve sıralamasını bilme ihtiyacı doğmuştur (38).

On dokuzuncu yüzyıl ikinci yarısı ve 1900'lerin başında İngiltere ve ABD'de sınavlar, üç farklı amaca yönelik kullanımlarıyla yaygınlık kazanmıştır. İşe alınma, yöneticilik pozisyonlarına getirilme ve üniversite eğitimine girişi belirlemek amacıyla yapılan "seçme sınavları"; okullarda verilecek eğitimden önce öğrencilerin düzeylerini belirlemek amacıyla yapılan "yerleştirme sınavları"; sınıf geçme, mezuniyet ve yeterliğe karar vermek amacıyla yapılan "sertifikasyon sınavları", farklı amaçlara hizmet eden sıralamaya ve nota dayalı sınavlar olarak kullanılmaya başlanmıştır (22,23,29,33,39). Aynı dönemde tıp, hukuk, mühendislik gibi mesleklerin edinilmesi ve uygulanmasını düzenleyen çok sayıda meslek örgütü kurulmuş ve bu alanlarda yeterlik kurulları tarafından yapılan sınavlar önemli bir yaptırım gücü kazanmıştır (29,32,40). National Board of Medical Examiners 1915 yılında kurulmuş, ilk yeterlik sınavını 1916'da gerçekleştirmiştir (41). Tıp fakültesine başvuran öğrencilerin seçimi aşamasında kullanılmak üzere, "Scholastic Aptitude Test for Medical Schools" adıyla Association of American Medical Colleges tarafından hazırlanan test, 1930 yılında uygulanmaya başlanmıştır (40).

Hekimlik, mühendislik gibi kendi alanlarına ait bilimsel bilgi birikimine sahip mesleklerde ehliyetlendirme amacıyla teorik yazılı sınavlar, okullarda da teorik bilgiyi kapsayan yazılı sınavların önem kazanmasına yol açmış; yazılı sınavların yüksek statülü meslekler ile ilişkilendirilmesi nedeniyle sınavın yazılı olması, sınavın kendisine itibar kazandıran bir özellik haline gelmiştir (10).

Yanıtlayan ve değerlendiren için tek ortak cevabı olacak biçimde soruların dikkatle düzenlenmesi, öğrenciler arasındaki farklılıkları ortaya koyabilecek ayırıcılık niteliği olan soruların kullanılması, notlandırmaya dayalı olarak sıralama yapılması gibi meritokratik düzenlemeler ile değerlendirmede nesnellik ve standardizasyon ön plana çıkmış; değerlendirme, belirleyiciliği yüksek önem taşıyan (high stake) bir nitelik kazanmıştır (23). Bu süreçte değerlendirme, öğrenme ile organik ilişkisini ve didaktik, eğitsel fonksiyonunu kaybetmiş, sosyal katmanlar arasındaki geçişin gerçekleşmesi açısından toplumun yapısal olarak yeniden organize edilmesinde kullanılan bir "teknoloji"ye dönüşmüştür (22,23,28,42). Bir önceki yüzyılda öğretimin içinde, öğrenme ve öğretmeden ayrı olarak görülmeyen değerlendirme, artık eğitim alanında ayrı bir süreç ve konu olarak ele alınmaya başlanmıştır (36). Verimlilik temelinde belirlenen hedef davranışlara göre eğitim içeriğinin düzenlenmesi ve değerlendirmenin uzmanlık gerektiren teknik süreçlere dönüşmesiyle, eğitim programlarının geliştirilmesi ve öğrencilerin değerlendirilmesi profesyonel alanlar haline gelmiştir (23,37). Soru tipleri, değerlendirme yöntemleri, soru ve sınavların hazırlanmasına ilişkin teknik kural ve süreçler tanımlanmış, değerlendirme alanındaki temel kaynaklarda bu konulara ilişkin ayrı bölümlere yer verilmiş ve rehberler hazırlanmıştır (39). American Psychological Association (APA-1892), American Educational Research Association (AERA/NADER-1916), American Council on Education (ACE-1918), National Council on Measurement in Education (NCME-1938) gibi organizasyonlar, bu uygulamaların geliştirilmesine yönelik yapılan çalışmalar ve standartların belirlenmesinde söz sahibi kurumlar olmuştur (3,4,5). Brookhart'a göre (44), eğitimde ölçmenin (educational measurement) bir alan olarak gelişmesi sırasında sahip olduğu varsayımlar ve yapılan tercihler, kalıtıma dayalı zekâ kuramı, davranışçı öğrenme kuramı, "sosyal verimlilik" anlayışına dayalı eğitim programı ve psikometrik temelli "bilimsel ölçüm" düşüncesi tarafından biçimlendirilmiştir.

2. "Değerlendirme"den "Ölçme Değerlendirme"ye - Epistemolojik Bağlam (20. yy)

Aydınlanma Çağı ile 18. yüzyılda düşünsel temelleri atılan, Fransız Devrimi ile siyasi ve hukuki üst yapısı, Sanayi Devrimi ile sosyo-ekonomik alt yapısı biçimlenen modernleşme hareketi, 19. yüzyılın tamamı ve 20. yüzyılın ilk yarısı boyunca, toplumsal, siyasal, ekonomik ve kültürel yaşamın belirleyicisi olmuştur. Modernite, sosyal ve siyasal bir değişim olmanın ötesinde, dönemin düşünme sistemini oluşturmuştur. Aydınlanma düşüncesinden hareketle, nesnel, evrensel ve tek bir gerçeklik bilgisinin akıl ve deney yoluyla elde edilebilir olduğu görüşü (9) çerçevesinde, bilim ve bilimsel alandaki gelişmeler, sanayileşme yolundaki modern toplum için itici bir güç haline gelmiş ve "bilimsellik" iddiası önem kazanmıştır. Modern çağın egemen paradigması olan pozitivizm, bizim dışımızdaki gerçekliğin, akıl yolu ile bilim tarafından doğru ölçüm ile araştırılıp tanımlanabileceğini ve bilimsel olmayan bilginin "bilgi" kabul edilemeyeceğini savunmuştur (45). Pozitivizm, "bilginin" ne olduğu, kapsamı ve kaynağına ilişkin bu ön kabullere dayalı olarak, genellenebilir bilginin nasıl elde edilebileceğine yönelik metodolojiyi, uygun örneklem seçimi, gözlemlenebilir ve ölçülebilir değişkenler, nesnellik, güvenirlik, tekrar edilebilirlik kavramları ile tanımlamıştır.

Önemle vurgulayıp ortaya koyduğu metodolojik farklılık üzerinden 'kesin' ve 'gerekli' olan "hard science" ile 'belirsiz' ve 'gereksiz' olan "soft science" ayrımı yapan pozitivist bakış (46), sosyal bilimler ve davranış bilimlerinin çalışma alanı ve yöntemleri için de belirleyici rol oynamıştır (9,46).

Pozitivist ve rasyonel paradigmadan hareketle, sosyal olayların, insanın duygu, düşünce ve davranışlarının, doğa/fen bilimlerinin kullandığı yöntemlerle araştırılıp açıklanması, antropoloji, psikoloji, sosyoloji, ekonomi ve eğitim gibi alanlarda bilimsel düşünme ve araştırmayı etkileyip değiştirmiştir. Doğa/fen bilimlerinin "ölçüm" temelli, objektif ve standart biçimde yapılan deney ve gözlemler sonucu elde edilen sayısal verilere dayalı araştırma metodolojisi, sosyal bilim ve davranış bilimlerinin de temel metodolojisi haline gelmiş; "ölçme", bilimsel bilginin belirleyici özelliği olarak görülmüştür (9).

Sosyal yaşamın, zihinsel yapı ve süreçlerin sayısallaştırılarak incelenmesi ve "ölçmenin" bir hedef haline gelmesiyle, sosyal bilimler ve davranış bilimleri alanında "ölçme" teknikleri, istatistik yöntemler gelişmiş; pozitivist paradigmaya uygun yöntemlerin kullanılması, sosyal bilimlerin ve özellikle psikolojinin "bilimsel meşruiyet" kazanmasını sağlamıştır (9,38,47,48,49).

Modern psikolojide deneysel ve kantitatif metodolojinin kullanılması, 1800'lerin ikinci yarısında Alman psikolog Fechner'in duyuların yoğunluk ya da şiddetini sayılaştımaya yönelik psikofizik alanında yürüttüğü çalışmalarla başlamıştır (6). Daha önceden meta-psikoloji olarak normatif alanda tanımlanan psikoloji, davranışçı psikolojinin kurucularından sayılan Watson'un 1913 yılındaki ifadesiyle, "davranışçı kuramlar ile nesnel ve deneysel çalışmalar yapan fen bilimlerinin bir dalı olmuştur" (50). Doğrudan gözlenemeyen örtük zihinsel özelliklere (constructs-latent traits) ilişkin yapılan tanımlamalar ve bu özelliklerin aralarında kurulan ilişkiler (nomologic network) çerçevesinde geliştirilen kuramsal bir yapının ölçülmesi düşüncesi ve girişimleri, 20. yüzyılın başlarında "psikometri" adıyla ayrı bir profesyonel alan açmıştır (9). Ölçüm sonuçlarının bireyler/gruplar arasındaki farklılıklara göre yorumlanmasından hareketle, "klasik test teorisi" gibi "ölçme kuramları" ve istatistiksel yöntemler geliştirilmeye başlanmıştır (9,37).

Bilimsel yöntemlerin kullanılmasına yönelik ortaya çıkan metodolojik yaklaşım değişikliğinin yanı sıra, Darwin'in 1859'da "Türlerin Kökeni" adıyla yayınlanan çalışması, psikoloji alanında çalışan araştırmacıların ilgi alanları üzerinde de etkili olmuştur. Canlılar arasında türlerin değişerek başka türlerin ortaya çıkışında "işlevlerin" farklılaşmasının rolünü ortaya koyan ve bu değişim ile çevreye adaptasyon ve doğal seleksiyonun ilişkisini savunan Darwin'in bu çalışması, psikoloji alanında çalışan araştırmacıların bireyler arası farklılıkları ve bunların işlevlerini incelemeye yönelmelerinde rol oynamıştır (32,51-54). Her şeyin ölçülebileceğine inanan İngiliz araştırmacı Galton (1822-1911), antropometrik ve psikometrik ölçümler kullanarak bireylerin fiziksel özellikleri ve davranışlarını incelemiş, reaksiyon zamanı ve duysal farklılıklara göre yaptığı zekâ ölçümleri ile zekâ üzerine sonradan yapılacak olan çalışmalar kadar, "ölçmenin" psikoloji alanında kullanılmasına da öncülük etmiştir (55-58). Bireyler arasındaki farklılıkları incelemek için normal dağılım eğirisini ilk kullanan araştırmacı Galton olmuştur (58). Kalıtsal özellikler temelinde bireyler arasındaki farklar üzerinde çalışan ve diferansiyel psikolojinin kurucusu olan Galton'un kullandığı istatistik tekniğini değiştirip geliştiren Pearson ve Spearman gibi araştırmacılar, psikometrinin öncü isimleri arasında yer almışlardır; Pearson'un Ki-Kare testi (1900), Spearman'ın özellikle psikolojik verilere uygulayarak kullandığı korelasyona dayalı teknikler, sıra

korelasyon katsayısı ve faktör analizi (1904), psikolojik ölçme ve istatistik çalışmalarında geniş bir kullanım alanı bulmuştur (32,33,39,57,59). Psikometri, zekânın ölçümü ve kişilik testleri üzerine yapılan çalışmalarla beraber gelişmiştir (28,33,58,60).

Michell'e göre (6,60), herhangi bir konuda bilimsel olma iddiasının ölçmeyi gerektirdiği görüşü, ölçme alanı bulmak için modern psikoloji üzerinde bir baskı oluşturmuştur. Psikoloji alanında, ölçme yoluyla sayısallaştırmayı "sayısal zorunluluk" (quantitative imperative) olarak adlandıran Michell (6,60), psikolojik ve zihinsel özelliklerin niceliksel olduğu hipotezine dayanan psikometrinin, ölçek ve testler aracılığı ile zihinsel yapıların, kişilik özelliklerinin, sosyal tutum ve tavırların ölçülebileceği iddiasına işaret edip bu iddiayı eleştirmiştir. Michell, herhangi bir özelliğin ölçümünden önce, bu özelliğin niceliksel olduğuna dair hipotezin test edilmesi, özelliğin niceliksel olup olmadığının, dolayısıyla ölçülüp ölçülemeyeceğinin sorgulanması gerektiğini belirtmiş ve tüm özelliklerin niceliksel olduğu inancının temelsizliğine dikkat çekmiştir (6). Psikolojinin, operasyonalizmin etkisinde kaldığını öne süren Michell'in bu görüşü, farklı yazarlar tarafından da savunulmuştur (6,32,48,49,60).

Fizik alanında Bridgman tarafından 1927 yılında öne sürülen operasyonalizm, bilimsel kavramlar ve bu kavramlarla bağlantılı gözlemler arasındaki ilişkiye dair bir kuramdır; ölçülen özelliği, onu ölçmek için kullanılan işlemle tanımlayan operasyonalizme göre, teorik bir kavram, onu ölçmek için kullanılan araç ve işlemlerle eşdeğer bir anlam kazanır (32,48,61,62). Uzunluk kavramının, uzunluğu ölçtüğümüz metrik sistemden bağımsız düşünülmemesi, ölçme işlemini temel alan operasyonalizm için en sık verilen örnektir. Fiziksel nesneler ve olaylar için fizik alanında karşılığını bulan operasyonalizmin psikoloji tarafından benimsenmesinin, felsefi ve bilimsel olmaktan çok, politik bir yaklaşım olduğu öne sürülmüştür (32,48). "Tanımlama yapmadan ölçmek" ile eş anlamlı görülen zekâ testleri, psikolojide operasyonalizmin ilk örnekleri arasında sayılmıştır (32). Davranışçı psikolojinin ilk temsilcilerinden olan Edward Boring, 1923 yılında New Republic dergisinde yayınlanan "Intelligence as the Tests Test It" başlığıyla yayınladığı makalesinde, "yeni bir bilimsel gözlem bizi bu tanımdan uzaklaştırıncaya kadar, zekâ, testlerin ölçtüğü şeydir" tanımını yapmıştır (6). Benzer biçimde, Army Alpha zekâ testinin geliştiricileri arasında bulunan Goddard, "zekânın ne olduğunu bilmiyoruz, bilgininin ne olduğunu bilip bilmeyeceğimiz de şüpheli" ifadesini kullanmıştır (39).

Psikometrinin ve ölçme kuramlarının gelişmesi, eğitim alanında değerlendirme yaklaşımını köklü biçimde etkilemiş, geniş ölçekli, standart ve nesnel testler, tüm eğitim kademeleri için belirleyici olurken değerlendirme kavramını da değiştirmiştir. Değerlendirmeye ilişkin söylem ve terminoloji, "ölçme" temelinde şekillenmeye başlamıştır. Eğitim alanında ölçme temelli değerlendirme yaklaşımı, örtük özelliği/yapıyı (construct) ölçmek, bu ölçümü popülasyonun normlarına göre karşılaştırıp bir ölçekte açıklamak üzere psikoloji alanında geliştirilen ölçme anlayışı ve psikometrik kuramların üzerinde şekillenmiştir (10,36,51,63,64); nesnellik, standardizasyon ve güvenirlik "ölçüm" için temel vurgulanan konular olmuştur (10,28,65,66). Klasik test teorisi, 20. yüzyıl boyunca tüm test uygulamalarında temel alınmıştır (58). Ölçülmek istenilen değişkene ait gerçek değeri, gerçek puan olarak adlandıran klasik test teorisi, ölçme yoluyla doğrudan elde edilemeyen gerçek puanın, bazı varsayımlarla gözlenen puanlardan kestirileceği düşüncesine dayanmıştır (5). Borsboom'a göre, gözlemlenemeyen bir özelliğe ilişkin "gözlenen puan" ve "gerçek puan" ilişkisine dayalı klasik test teorisi ile operasyonalist düşünce birbirini karşılıklı olarak besleyen düşünme biçimleridir

(49).

Pozitivizmin egemen olduğu ve ölçmenin "bilimselliğin" ön şartı olduğu düşüncesi, kurumlar ve toplum üzerinde büyük etki göstermiş, 20. yüzyıl başında test kullanımının yaygınlığı ve sıklığının tüm dünyada hızla artmasıyla, bu testlerin "teknik bir iş" olarak hazırlanması ve uygulanması, özellikle ABD'de, yüksek kazançlı bir ticari sektörün ortaya çıkmasına yol açmıştır (33,47,58). Michell'e göre, psikoloji alanında yürütülen bilimsel çalışmalar ve verilen hizmetlerde "ölçme söyleminin" kullanılmasının, profesyonel alan olarak psikolojiye getirdiği ekonomik ve sosyal avantajlar tarihi kayıtlarda izlenebilmektedir (6).

3. Üç Farklı Alanda Test Uygulamaları (Test Çağı)

Delandshere'e göre (23), psikolojik testlerin ve psikoloji alanındaki ölçmenin 19. yüzyıl sonu ve 20. yüzyıl başında gelişmesi bir tesadüf değildir. Seri üretim için yaygın kitlesel eğitime ve yeni ekonomik-politik düzenin ilerlemesini sağlayacak en uygun ve yetenekli bireylerin "seçilebilmesine" duyulan ihtiyaç bağlamında, standardize testlerle ilgili psikometrik gelişmelerin olması, dönemin ekonomik ve sosyopolitik koşulları ile uyumludur (23). Bu dönemde, sosyal verimlilik ve "bilimsel yönetim" ilkelerine uygun düzenlenen eğitim programları, ilişkilendirici (associationist) ve davranışçı öğrenme kuramları ile bireysel farklılıkları açıklayan kalıtsal özellikleri dikkate almış; bu kuramlardan hareketle başarı ve yeteneğin "bilimsel ölçümü" büyük önem kazanmıştır (36). Thorndike ve Hagen, 1915-1930 arasını, yeni testlerin tavşanlar gibi çoğaldığı "patlama dönemi" olarak adlandırmışlardır (32). 1928 yılında Walter Monroe, ABD'de 1300 farklı standardize testin kullanımda olduğunu ve bir yıl içinde 30.000.000 kadar test sonucunun elde edildiğini belirtmiş; karşılaştırmalı bir tahminle 1944 yılında, yaklaşık 5000 ayrı testin 60 milyona yaklaşan kullanımı olacağından söz etmiştir (43). Test uygulamalarına ilişkin varılan nokta ve sonraki gelişmeler için oluşturulan temel açısından 1927 yılı sembol yıl olarak görülmüştür (58).

Zekâ, yetenek, kişilik özellikleri ve başarıyı ölçmek için teknik bir uzmanlık ile standart ve nesnel biçimde hazırlanıp uygulanan ve analiz edilen testler, psikoloji, çalışma dünyası ve eğitim olmak üzere üç alanda yaygın biçimde kullanılmaya başlanmıştır (32,33,38,55). Psikoloji alanında zekâ testleri, işe uygunluğun değerlendirilmesinde yetenek testleri ve eğitim alanında başarı testleri belirleyici uygulamalar haline gelmiş; "değerlendirme" konusuna ilişkin söylem, geçerlik tartışmaları ve değerlendirme kavramsallaştırması bu uygulamalar bağlamında biçimlenmiştir (32-34,38). Bu testlerin nitelikleri, uygulanma amaçları, kullanım biçimleri, etki ve sonuçları, geçerlik üzerine yapılan tartışmaların içeriği kadar kronolojisini de belirlemiştir.

Bu testlerin bağlamı kendi uygulamalarımızdan tamamen farklı olsa da, değerlendirme alanında günümüzde kullanılan "bilimsel söylem" ve değerlendirme anlayışını belirleyen psikometrinin tarihsel mirasını anlamak açısından önemlidir. Baird'e göre, öğrenmeye ilişkin farklı felsefe ve kuramlar benimsenmiş olmasına karşın, tarihsel miras, uygulamada değerlendirme sistemlerini zımnen şekillendirmektedir (67).

3.1. Psikoloji alanı ve zekâ testleri

Galton ile birlikte Londra'da bireysel farklılıklar üzerine yaptığı çalışmaları, ABD'ye döndükten sonra sürdüren Cattell, 1890 yılında yayınladığı "Mental Tests and Measurements" adlı ünlü makalesinde "mental test" kavramını ilk kullanan araştırmacı olmuştur (55,58). Galton'un antropometrik yaklaşımından biraz daha farklı olarak, Cattell, zekâyı fizyolojik ölçümler ile ilişkilendirmiştir (33,55). Aynı dönemde üniversite öğrencileri üzerinde yapılan

çalışmalarda, Catell'in yaklaşımı temelinde geliştirilen testlerden elde edilen sonuçlar ile akademik performans arasındaki korelasyon incelenmiş ve bu testlerin beklendiği şekilde akademik başarıyı ön görmedikleri ortaya konarak zihinsel özelliklerin ölçümü için fizyolojik yaklaşımın uygun olmadığı görüşleri öne sürülmüştür (33,55).

Yirminci yüzyılın ilk yıllarında yapılan bu çalışmalara paralel olarak Alfred Binet, zekânın sadece reaksiyon zamanı gibi temel duysal ve fizyolojik süreçlerle açıklanamayacağını, daha ileri psikolojik süreçlerin zekâ ölçümü için gerekli olduğunu belirtmiştir (55). Bu düşünceden hareketle yürüttüğü çalışmalarını, okullarda öğrenme güçlüğü çeken çocukların belirlenmesi ve uygun eğitim alabilmelerine yönelik dönemin eğitim uygulamaları doğrultusunda sürdüren Binet, Herbert Simon ile birlikte zihinsel gelişim temelinde öğrencileri sınıflandırabilecek Binet-Simon ölçeğini geliştirmiştir (33,55).

Eğitimleri sırasında akranlarından geride kalan çocuklarla ilgili bir komitede çalışmak üzere Fransız Eğitim Bakanlığı tarafından görevlendirilen Binet'nin, Simon ile beraber geliştirdiği ölçek, mental olarak yetersiz öğrencilerin diğer öğrencilerden ayrılarak farklı sınıflara yerleştirilmesi için ilk kez 1905 yılında Fransa'da kullanılmıştır (32-34,55). Öğretmenleri tarafından başarılı ve başarısız olarak değerlendirilen çocukları belirlemedeki ayırıcılık fonksiyonları, bu testi oluşturan maddelerin seçiminde dikkate alınan temel özellik olmuş (28,33); testte yer alan maddelerin düzenlenmesinde içerik değil, yaşa göre zorluk düzeyleri dikkate alınmıştır (55,58). Binet, geliştirdikleri ölçeğin hiçbir şekilde zekânın mutlak bir ölçüsü olamayacağını belirtmiştir (6,32,33). Üçüncü revizyonu 1911 yılında yapılmış olan Binet-Simon ölçeğinin "ölçme" amacı taşımadığı, sadece yaşa göre "sınıflandırma" amacıyla geliştirildiği özellikle vurgulanmıştır (55).

Zekâya ilişkin ilk sayısallaştırma, Stanford Üniversitesi'nde psikoloji profesörü olarak çalışan ve 1916 yılında Binet-Simon ölçeğini revize edip Stanford-Binet ölçeğini hazırlayan Terman tarafından yapılmıştır; test maddelerinin seçiminde iç tutarlılığı temel alan (58) Terman, bu ölçekte yaşa göre oluşturduğu katsayıya "Intelligence Quotient – IQ" adını vermiştir (33,37,55,56). Stanford-Binet ölçeği, zekânın ölçümünde uzun yıllar standart olarak kabul edilmiş, sonradan geliştirilen testlerin neredeyse tamamının geçerliği, bu ölçekle korelasyonlarına göre incelenmiştir; Stanford-Binet ölçeğinin etkisi ve yaygın kullanımı 1960'lı yıllara kadar devam etmiştir (55,58). Öjenik görüşleri çerçevesinde, zekâyı kalıtımla ilişkisi yönünden ele alan Terman'ın bu görüşlerini o dönemde de eleştirenler olmasına karşın, Kaestle'a göre (33), zekâ alanındaki çalışmaları ve geliştirdiği ölçek ve testlerle Terman, 1923 yılında APA'nın başkanlığına gelecek saygınlık ve şöhretle beraber yüksek maddi kazanç elde etmiştir.

Zekâ geriliği olan çocukların eğitimi ve mental retardasyon üzerine araştırmalar yapan ve Binet-Simon ölçeğini ABD'de yürüttüğü çalışmalarda tanısal amaçlı kullanarak "moron" teriminden ilk söz eden psikolog Goddard olmuştur (55). Goddard'ın 1912 yılında yayınlandığı "The Kallikak Family: A Study in the Heredity of Feeble-Mindedness" adlı kitapla beraber öjenik politikalar gündeme gelmiştir (33). Goddard, geri kalmış bölgelerde ve küçük göçmen grupları üzerinde yürüttüğü çalışmalarla elde ettiği verileri kalıtsal özelliklere bağlı olarak yorumlamış ve göçmenlerin zekâ ortalamalarının düşük olduğu sonucuna varmıştır (33,55).

Terman ve Goddard'ın sahip oldukları saygınlık ve etki gücü ile beraber öne sürdükleri görüş ve değerlendirmeleri, ABD'de, ulusal zekâ düzeyinin düşeceği gibi endişe ve memnuniyetsizliğe yol açarken (33,34), kalıtım yoluyla daha iyi nesillerin geliştirilmesine yönelik olarak kısırlaştırma, göçmen kabulünün sınırlandırılması gibi öjenik politikalar ABD'de

tartışılmaya başlanmıştır (33,55). 1900'lerin başında çok etkili bir isim olmasına karşın, psikolojik testlerin tarihine ilişkin temel kitaplarda Goddard'a fazla yer verilmediğini belirten Gregory (55), dönemin ideolojisi ve sosyal normları çerçevesinde psikolojik test ve ölçümlerin en etkili isimler tarafından bile yanlış kullanıldıklarına işaret etmiştir.

Yirminci yüzyılın ilk yıllarında zekâ ve zekâ testleri üzerine farklı çalışma ve tartışmalar devam ederken, bu testlerin bireylere tek tek uygulanmasının zaman alıcı olması nedeniyle, çok sayıda kişiye aynı anda uygulanabilecek testlerin geliştirilmesi için araştırmalar başlamıştır. Terman'ın Stanford Üniversitesi'nde öğrencisi olan Arthur Otis'in, doktora çalışması kapsamında çoktan seçmeli test tekniği ve çabuk skorlama yöntemini kullanarak geliştirdiği grup halinde uygulanabilecek ilk zekâ testi (Group Intelligence Scale) üzerindeki araştırmaları, I. Dünya Savaşı ile beraber hız kazanmıştır (32,56). Zekâ testlerinin ilk geniş ölçekli uygulaması, Otis'in geliştirdiği teknik ve yöntemi model alan diğer bir çalışmayla, I. Dünya Savaşı sırasında gerçekleştirilmiştir (32). Bu çalışma, savaş hazırlıklarını desteklemek üzere, Harvard Üniversitesi'nden Robert Yerkes'in o tarihlerde başkanı olduğu American Psychological Association (APA) tarafından 1917'de oluşturulan bir komite tarafından yürütmüştür (68). Yine Yerkes'in başkanlığında, Boring, Terman, Goddard, Thorndike, Thurstone, Otis gibi dönemin en etkili akademisyen ve araştırmacı psikologlarından oluşan bu komite, ordudaki görevlere seçme ve yerleştirmede kullanılmak üzere Army Alpha ve Beta adı verilen, gruplara uygulanabilecek zekâ testlerini hazırlamıştır (32,33,55,56,68). Ordudaki üst rütbeli subay ve yöneticilerin itiraz ve eleştirilerine karşın (33,55), Yerkes'in ABD yönetimi ve orduyu ikna etmesiyle (33,56), sayıları 100'ü aşan sivil psikolog tarafından (33), 1918 itibariyle askere başvuran yaklaşık 2 milyon kişiye zihinsel kapasitelerinin ölçümü için bu testler uygulanmıştır (32,33,55,68). Bu komitenin çalışmaları, psikoloji alanında, APA'nın içinde "Division of Military Psychology" adıyla, 1946 yılında yer alacak olan, yeni disiplinin ortaya çıkıp kurumsallaşmasının başlangıç noktasını oluşturmuştur (56,69). Bu testler aracılığıyla yapılan ölçümlerin hataları sonradan gösterilmiş olsa da (55), Army Alpha ve Beta testleri ve yol açtıkları tartışmaların, bu yıllarda psikolojinin tanınması yolunda Freud'un çalışmalarından bile daha fazla etkili olduğu öne sürülmüştür (56). Bu testler, 20. yüzyıl boyunca sonraki test uygulamaları için model olmuştur (37,58).

Army Alpha testi, psikologlar için test geliştirme ve psikometri konusunda zengin bir deneyim sağlamış, farklı değişkenler arasında binlerce korelasyon katsayısı hesaplanarak veriler analiz edilmiş ve test/ölçek geliştirme, kısa sürede "sanat olmaktan bilim olmaya terfi" etmiştir (55).

I. Dünya Savaşı yıllarındaki bu uygulamalar, test geliştirmenin, içeriğin dikkate alındığı "logical" bir süreç olmaktan çok, "ampirik" bir işlem olduğu yönündeki algıyı yaygın bir biçimde güçlendirmiştir (32). Testler, kabul edilen bir kriterle ilgili en iyi öngörüyü sağlayabilecek biçimde hazırlanmaya çalışılmış ve "ölçüt" (criterion) kavramı genel kullanıma girmiştir (32).

Army Alpha'nın geniş çapta uygulanması, "mental testing" teknolojisinin tanınırlık ve saygınlığını artırmış (32,33,55,69), savaş sonrasında, eğitim alanında ve çalışma hayatında işe alımlarda kullanılmak üzere bu yöntem hızla ve yaygın biçimde farklı alanlara transfer edilmiştir (32,37,38). Otis'in Group Intelligence Scale adını verdiği test, Army Alpha tarafından model alınmasını takiben, 1918 yılında gruplara uygulanan ilk ticari test olmuş, bunu Terman'ın Army Alpha üzerinde az sayıda değişiklik yaparak hazırladığı Terman Group Test of Mental Ability (1920) adlı test izlemiştir (56). Avrupa ve ABD'de benzer

çalışmalarla geliştirilip ticarileşen "mental testler", toplumsal yapının işleyişini sağlamak amacıyla bireylerin seçim ve sıralamasını gerçekleştirmek için yaygın biçimde kullanılmış (34), 1922 yılına kadar, ABD'de yılda yaklaşık 3 milyon çocuğa uygulanmıştır (47).

Zekânın kalıtıma dayalı sabit, değişmez bir özellik olduğu görüşüne karşı, 1940'lı yılların ortalarından sonra, sosyal çevre ve koşulların zekâ üzerindeki etkilerini dikkate alan görüşler ortaya çıkmış ve zekâ testleri sorgulanmaya başlanmıştır (7). Yirminci yüzyıl başlarındaki okul kavramsallaştırması içinden geliştirilen zekâ testlerinin yaygın kullanımını izleyen 1950'li ve 60'lı yıllarda yapılan çalışmalar, testlerin yanlılık (bias) içerdiğini, yapılan ölçümlerin yanlışlığını ortaya koymuş; zekâ testleri, sosyal eşitsizliklerin meşrulaştırılarak sürdürülmesinde rol oynadıkları gerekçesiyle eleştirilmiştir (28,33,37,55).

3.2. Çalışma hayatı ve yetenek testleri

1911 yılında yayınlandığı "The Principles of Scientific Management" adlı inceleme ile fen bilimlerin metodolojisinin işletme alanında kullanılmasını savunan Frederick Taylor, "bilimsel yönetim" (scientific management) hareketinin kurucusu sayılmaktadır (38,70). Çalışma yaşamında işlerin analiz edilmesi, bu analize göre işlerin gerektirdiği nitelik ve sürenin belirlenmesiyle üretimde etkinlik ve verimliliğin artacağını öne süren Taylor'un görüşleri, dönemin temel özelliklerinin, sanayileşme, kapitalizme inanç, ilerlemecilik ve pozitivizm olarak sayıldığı 20. yüzyıl başlarında, çalışma hayatı, üretim yönetimi ve eğitim alanında etkisini göstermiştir (37,38,70). İmalat sanayinin hızla geliştiği, özel sektör ve devlete ait işletmelerin sayısının katlanarak arttığı bu dönemde, işe başvuran çok sayıda kişi arasında vasıflı, vasıfsız insan gücü gerektiren işlere göre "bilimsel yönetim" ilkeleri çerçevesinde seçim yapılabilmesini mümkün kılan "teknoloji" ye ilişkin talep gelişmiştir (68). Aynı yıllarda, insanların zihinsel ve fiziksel özelliklerini ayırabilecek testler üzerinde araştırmalar yapan Galton, Binet, Cattell gibi isimlerin çalışmaları, bu talebi karşılamaya yönelik girişimler için yol gösterici olmuştur (68). "Verimlilik" hareketi ve Taylorizm, 1921 yılında İngiltere ve ABD'de "endüstriyel psikoloji" adıyla yeni bir alanın ortaya çıkmasında etkili olmuş (68,71), personel/çalışan seçimi bu alanın temel çalışma konuları arasında yer almıştır (38,53). Geliştirdiği psikometrik yöntem ve testlerle bu alana katkıda bulunan Thurstone, 1923 yılında ABD'de devlete bağlı Kamu Çalışanları İdaresi Bürosu'nun yöneticiliğine atanmış ve işe alımlarda rekabetçi olduğu kadar sistematik bir seçme için psikometrik değerlendirmeyi devlet kurumlarında uygulamaya koyup, kurumsallaştırmıştır (56,68). İşe uygunluğun değerlendirilmesinde kullanılan yetenek, tutum ve zekâ testleri, 1910'lu yılların ortalarından sonra İngiltere ve ABD'de özel sektör ve devlet kurumlarında işe alım süreçlerinin bürokratik ve rutin bir parçası haline gelmiştir (53,68).

On dokuzuncu yüzyılın son çeyreğinde Kuzey ve Batı Avrupa'dan göç almış olan ABD, 20. yüzyılın ilk on yılında Doğu ve Güney Avrupa'dan yeni bir göçmen akınına uğramış ve bu yıllarda göçmen sayısı, önceki döneme göre üç katına çıkmıştır (28,33). Doğu ve Güney Avrupa'dan gelen büyük sayıdaki göçmen nüfus nedeniyle iş gücü fazlasından kaygı duymaya başlayan işçi sendikalarının endişeleri, "bilimsel testlerin meşruiyeti" ile giderilmiş; iş analizi, işe uygunluğun değerlendirilmesi için yapılan testler ve psikometrik yöntemlerin kullanımı giderek artmıştır (53,68). Bu testlerde göçmenler ve yoksulların daha düşük puanlar almaları, az ücretle, vasıfsız işlere yerleştirilmelerine yol açmıştır (33,68). Yapılan testler, çalışma verimini gösterecek niteliğe sahip olmaktan uzaklaşmış, kısa zamanda önyargılara dayalı, ayrımcı, fırsat eşitliği tanımayan bürokratik uygulamalara dönüşmüştür (33). Kullanılmakta

olan testlerin, işe alımlarda birden fazla özelliği ölçmek için uygun olmadığını, farklı yetenek ve özelliklerin ölçülmesi için test takımlarından oluşan ölçeklerin gerekli olduğunu düşünen Thurstone, faktör analizi üzerinde çalışarak kişilik özelliklerini ölçmek amacıyla 1930'lu yıllarda çok faktörlü yeni bir model geliştirmiştir (55-58).

I. Dünya Savaşı gibi II. Dünya Savaşı da, profesyonel alan olarak genel anlamda psikolojinin uygulamalı katkılarının tanınıp kabul edilmesi ve bu alanın hizmetlerine talebin artmasında önemli rol oynamıştır (8,9). II. Dünya Savaşı sırasında "Office of Strategic Services" (OSS) adıyla kurulan ve çok sayıda uzman psikoloğun çalıştığı birim, yöneticilerin ve liderlerin seçimi ve eğitimi üzerinde yeni bir sistem ortaya koymuştur (38,56). Kağıt-kalem testlerinin dışında, durumsal olaylara verilen tepki ve performansın gözlemciler tarafından değerlendirildiği ve bir hafta boyunca süren testler uygulanmış, 13.000 kişiye ait puanlamalar ve biyografik bilgileri içeren veri elde edilmiş ve çoklu korelasyon, faktör analizi gibi teknikler kullanılarak bu veriler üzerinde çalışılmıştır (38,56,69). Bu çalışmalar, sonraki yıllarda psikoloji alanında kişilik testlerine ilişkin araştırmalar için zemin oluşturmuştur (13,69). Kişilik testleriyle yapılan ölçüm ve yorumların yanlışlığı, 1950'li yıllarda geçerlik tartışmalarının merkezinde yer almış ve yapı geçerliği kavramı ilk kez, kişilik testlerinin geliştirilmesi ve kullanılmasına ilişkin sorunlara yönelik tartışmalar bağlamında ortaya çıkmıştır (17).

Avrupa ve ABD'de meritokratik anlayışla başlayan ve psikometrik ilkelere göre düzenlenen seçme ve yerleştirme sistemi "fırsat eşitliği" bağlamında eleştirildiği gibi, iş performansını öngörmedeki yetersizliği ve çalışma hayatının farklı dinamiklerini göz ardı etmesi nedeniyle de 1950'lerden sonra sorgulanmaya başlanmıştır (25,33,68). Psikometrinin "bilimsel yöntemleri" ile çalışma dünyasının gerçekleri arasındaki uyumsuzluk, personel seçiminde kuramsal ve metodolojik yeni arayışların ortaya çıkmasına yol açmış, paradigma değişikliğinin tartışılmaya başlanması ile beraber endüstriyel psikoloji disiplini içinde İnsan Kaynakları Yönetimi (Human Resources Management) alanı ortaya çıkmıştır (68).

3.3. Eğitim alanı, seçme-yerleştirme ve başarı testleri

Psikometri temelinde ölçme kuramları çerçevesinde hazırlanan eğitim amaçlı testler, seçme-yerleştirme ve eğitim hedeflerinin değerlendirilmesi gibi iki farklı amaca hizmet etmek üzere farklı teknik özellikleriyle 20. yüzyıl başından itibaren kullanılmaya başlanmıştır. Seçme amaçlı testler, bireyler arası zekâ farklılıkları üzerine yapılan çalışmalarla paralellik göstermiş ve eğitime başlamadan önce, öğrenmedeki başarıyı öngörme temelinde, bireyler arasındaki yetenek ve kapasite farklılıklarına göre seçilip seçilmeyeceklerine ait karar verme sürecine hizmet etmiştir (34). Eğitim hedefleri açısından başarıyı ölçmeye yönelik testler ise okullar için toplumsal sorumluluk ve hesap verebilirlik bağlamında ortaya çıkan politik ve sosyal talebe cevap vermek üzere kullanılmışlardır (34).

1910'ların sonunda çok sık ve yaygın olarak kullanılmaya başlanan zekâ testleriyle ilgili eğitim alanında en bilinen örneklerden birisi, ABD'de "Ulusal Araştırma Konseyi"nin (National Research Council)" "Ulusal Zekâ Testi" hazırlama girişimi olmuştur; aralarında Yerkes, Cattell ve Thorndike'ın bulunduğu beş seçkin psikolog, özellikle okullarda kullanılmak üzere bir test hazırlamak için 1919 yılında bir araya gelmişler ve 1920 yılında tamamladıkları bu test 200.000 kopya olarak satılmıştır (33).

ABD'de 1926 yılında çıkarılan zorunlu eğitim yasası ile ilk ve orta öğretime başvuran öğrenci sayısı 1910 yılına göre dört kat artmıştır (10). Zorunlu eğitim yasası ve göçmen nüfusun artışı ile birlikte farklı düzeyde kabiliyet ve kapasiteye

sahip çocukların aynı sınıfta yer alması sonucuna yol açmış; bu bağlamda, okullarda yetenek ve kapasitelerine göre öğrencileri gruplamak, farklı sınıf ve okullara yerleştirmek amacıyla testlerin kullanımı belirgin biçimde artmıştır (33,34). Kısa sürede eğitim uygulamalarının bir parçası haline gelip kurumsallaşan bu testler aracılığı ile en iyi eğitim sisteminin yaratılması amaçlanmış, seçme ve yerleştirme işlevi ile kullanılan testler, eğitim yönetiminin başarısında önemli bir faktör olarak görülmüştür (10,34). Liseye başvuran 14-17 yaş grubu öğrencilerin sayısındaki büyük artışa ve bu testlerin uygulanmasına paralel olarak farklı eğitim programlarına sahip olan ve üniversiteye geçiş imkânı tanınmayan meslek ve ticaret liseleri açılmıştır (33). ABD'de günümüzde de uygulaması devam eden, yüksekokul ya da üniversite başvurularının seçme ve kabul aşamasında puanlarının kurumlar tarafından farklı biçimde dikkate alındığı "Scholastic Assessment Test" (SAT), ilk kez 1926 yılında "Scholastic Aptitude Test" adıyla kullanılmaya başlanmıştır (33,55).

Kuzey ve Batı Avrupa'dan daha önce gelenlere göre etnik ve dini kimlikleri farklı ve daha yoksul olan Doğu ve Güney Avrupa'lı yeni göçmenler, ABD'de daha "yabancı" olarak algılanmış, yeni dalga göçmen akınının, Amerikan yaşam tarzını değiştirebileceği endişesi doğmuştur (33,34,55). Okul çağında yüksek sayıda göçmen çocuğunun olduğu bir ülkede, zorunlu eğitim yasası, bir anlamda "Amerikalılaştırmayı" (Americanization) okulun temel fonksiyonlarından birisi haline getirmiştir (34,37). Ülkenin demografik ve ekonomik değişimine cevap olarak yeni testlerin kullanımı hızla artarak önceki yüzyıldaki kullanım sıklık ve yaygınlığını geride bırakmıştır (33).

Zekâ geriliği ve özel eğitim ihtiyacı olan çocukları belirlemeyi amaçlayan zekâ ve yetenek testleri, Binet ve Simon tarafından bu çocuklara uygun eğitim fırsatı yaratma gibi humanistik bir motivasyon temelinde geliştirilmiştir (55). Ancak, Batı Avrupa ve ABD'de eğitim ve iş alanında seçme-yerleştirme amaçlı yaygın kullanımlarını takiben zekâ testleri, sosyal sınıf, ırk, etnisite ve cinsiyete dayalı ayrımcı uygulamaların araçları olmakla eleştirilmeye başlanmıştır (25,34,37,55,62). Zekâ ve yeteneği ölçmek üzere, eğitime başlamadan önce uygulanan bu testlerin sonucuna göre öğrenciler kategorize edilmiş, düşük kapasiteli olduğu belirlenen öğrenciler, farklı sınıflarda, akranlarına göre daha düşük yoğunluklu içerik ve görevlere göre düzenlenmiş programlara yerleştirilmişlerdir (28,34). Farklı eğitim programlarına devam eden öğrencilerin sonraki eğitim başarıları ve işe yerleşme fırsatları da akranlarına göre düşük olmuştur (33,34). Glaser ve Silver'a göre (34), böyle bir uygulama ile testler, erken dönemde çocukların kaderlerini belirlemeye başlamıştır. Sonraki akademik performansı ön görmek üzere kullanılan bu testlere göre eğitimde yapılan seçme ve yerleştirmelerin, orta ve üst sınıf, erkek ve beyazların lehine sosyal eşitsizliklerin meşrulaştırılması ve sürmesine zemin sağladığı belirtilmiştir (11,25,28,33,34,62,72).

Bu testler aracılığı ile yapılan ölçümlere göre belirlenen kategorizasyonların hatalı olduğu, düşük puan alanların sosyoekonomik statü, dil, kültür farklılıkları nedeniyle testte yer alan soruları yanıtlamakta güçlük çektikleri sonraki çalışmalarla belirlenmiş (25,28,33,34,37); bu çocuklar, seçme ve yerleştirme uygulamalarıyla sistemden dışlanıp, adil bir eğitim fırsatı alamadıkları gibi, stigmatize edilmişlerdir (28,34,62). Farklı yanlarıyla 1940'ların sonunda güçlü biçimde eleştirilmeye başlanan seçme ve yerleştirme testleri, psikometrik özellikleri ve eğitim politikaları açısından olduğu kadar ABD'de yargıya taşınan davalarla da 1950'li ve 60'lı yıllar boyunca kamuoyunda tartışılmıştır (28,32,33,73). Cronbach, bu testler ile ilgili eleştirilerin, "zamanın ruhuna" (zeitgeist), dönemin koşulları ve egemen inanışlarına göre farklı biçimlerde ele alındığını belirtmiştir (33).

Eğitim alanında kullanılan seçme ve yerleştirme testlerinin dışında, öğrenci başarısını değerlendirmek için eğiticiler ya da okullar tarafından yapılan sınavlar da, özellikle ABD ve İngiltere'de 20. yüzyıl başında standart ve nesnel testlere dönüşmeye başlamıştır (33,34). Eğitim içeriğinde yer alan konulara ve sınıf düzeylerine göre düzenlenmiş ilk standart aritmetik testi 1908 yılında ABD'de kullanılmıştır (33). 1917'de Alpha Army testinin hızlı skorlanabilen ve tam "objektif" modelinin kazandığı başarının da etkisiyle, 1920'lerde çoktan seçmeli sorular, başarı testlerinin çoğunda kullanılan soru tipi olmuş ve bu soru tipi, skorlama kolaylığı kadar özellikle objektif olma özelliğiyle de geniş ölçüde savunulup desteklenmiştir (33). Ticari sektörün kısa sürede sahiplendiği başarı testleri, ABD'de hızla yaygınlaşmış, I. Dünya Savaşı'na girildiğinde, 200'den fazla testin ilk ve orta öğretim kurumlarının kullanımı için pazarlanmış olduğu belirtilmiştir (34). Terman'ın Kelley ve Ruch ile birlikte hazırladığı ve 1923 yılında ilk kez basılan "Stanford Achievement Test" adlı başarı testi takımı, geliştirilmiş versiyonlarıyla ABD'de günümüzde de kullanılmaktadır (55). Aslında zekâ testleriyle aynı entellektüel ve kurumsal zemine sahip olan standart, nesnel özellikteki başarı testleri (33), belirli zaman aralıklarıyla geniş ölçekli olarak uygulanmaya başlanmış, özellikle ABD'de, öğretmen, kurum ve eğitim programlarının başarısının izlenmesi, okullar ve bölgeler arası öğrenci başarılarının karşılaştırılmasında kullanılarak kurumsallaşmıştır (34,37).

4. Değerlendirme anlayışı sorgulanıyor: "Ölçme"den "değerlendirme"ye

Yirminci yüzyılın ilk yarısında Batı Dünyasında her üç alanda da yaygın ve yerleşik uygulamalara dönüşen test kullanımı ile ilgili tecrübe üzerinden 1940'lı yıllardan itibaren testlere ilişkin farklı eleştiriler ortaya çıkmaya başlamıştır (32). İlk uzay aracı Sputnik'in 1957 yılında Sovyetler Birliği tarafından uzaya gönderilmesi, Amerikan eğitim sisteminin sorgulanmasında bir dönüm noktası olmuştur (30,33,34,51). II. Dünya Savaşı sonrası, bilim alanına yapılan büyük yatırım ve bilimsel araştırmalar için ayrılan geniş kaynaklara karşın, soğuk savaş yıllarında rakibi konumundaki ülkenin ulaştığı bilimsel gelişmişlik seviyesi ve yarattığı tehdit karşısında yaşanan hayal kırıklığı, ABD'de tüm eğitim sisteminin sorgulanmasına yol açmıştır (21,30,33,34). Bilimsel bilginin üreticisi ve kullanıcısı olarak görülen ve "teknik rasyonalite"ye dayalı olması beklenen meslek alanlarında, özellikle hekimlerin, mühendislerin ve öğretmenlerin mesleki uygulamalarında görülen yetersizlikler, ABD toplumunda memnuniyetsizlik yaratmıştır (30). Bu dönemde, her kademesiyle gözden geçirilen eğitim sistemi ile birlikte reform girişimleri başlamış, yaygın biçimde uygulanmakta olan testlerin psikometrik özellikleri, test sonuçlarına dayalı verilen kararlar ve testlerin etkileri farklı açılardan tartışılmıştır (33).

Testlerin psikometrik özelliklerine ilişkin sorgulama, geçerlik tartışmalarıyla bütünleşirken (32); ölçümlerin yanlı ve yanlış olduğu, ölçülen özellik üzerinde etkili olan sosyoekonomik durum, kültürel özellikler gibi faktörlerin testler yoluyla yapılan ölçümlerde dikkate alınmadığı yönündeki eleştiriler temelinde meritokratik uygulamalara duyulan güven ve testler, sosyal eşitsizlikler açısından da sorgulanmıştır (10,11,25,33,62,67). Meritokratik anlayış, sosyal sınıf farklılıklarını, gelir ve gücün farklı toplum kesimlerinde eşit olmayan dağılımını gözardı etmiştir (25). Meroe'ye göre (25), bireysel yetenek ve kapasitelerine göre herkese yeterli fırsat ve eşit şansın verilmesinden sonra ortaya çıkan eşitsizlikler, Sosyal Darwinist bir evrim gibi ele alınmıştır.

Testlerin kendileri ve yaygın kullanımlarıyla ilgili eleştirilen özelliklerin ötesinde, bu uygulamalar, yükseköğrenim dâhil eğitimin her kademesinde değerlendirmeye yönelik anlayışı biçimlendiren etkileri açısından da

ele alınmıştır. Bu uygulamalara dayalı olarak yapılan çalışmalar, tartışılan konu ve kavramlar, yayınlanan temel kaynaklar ve yürütülen geçerlik tartışmaları üzerinde ölçme kuramlarına dayalı değerlendirme kavramsallaştırması ve "bilimsel söylem" gelişmiş, toplumun, eğitim yönetiminin, eğitici ve öğrencilerin değerlendirme algı ve anlayışları biçimlenerek psikometrik değerlendirme anlayışı eğitim alanında egemen olmuştur (11,36,37,74). Geçmişin ya da psikometrinin mirası olarak adlandırılan bu etki öğrenme ve değerlendirme kültürünü şekillendirmiştir (10,23).

1950'lerden sonra, dünyada eğitim alanında reform hareketlerinin başlaması ve 1960'lı yıllarda bilişsel psikoloji alanındaki çalışmalara paralel bilişsel öğrenme kuramının gelişmesiyle testler, değerlendirme bağlamında öğrenmeyle ilişkileri açısından da sorgulanmaya başlanmış ve özellikle 20. yüzyılın sonlarına doğru tartışmalar "değerlendirme anlayışı" üzerinde yoğunlaşmıştır (23,28,33,34,51,74). Ölçüm temelli psikometrik değerlendirme anlayışının egemenliğine karşın, öğrencilerin gerçek performanslarına ilişkin bilginin değerlendirme yoluyla elde edilip edilemediği, değerlendirmenin öğrenme ve öğretme süreçleri üzerindeki olumlu olumsuz etkilerinin neler olduğu gibi sorular üzerinden değerlendirme anlayışında değişimin gerekliliği öne sürülmüştür (16,63,75-79). Gipps, değerlendirme için "teknolojiye" değil, değerlendirmenin amaç ve niteliğine ilişkin farklı ve radikal bir kavramsallaştırmaya ihtiyaç olduğunu belirtmiştir (10). Yirminci yüzyılın ikinci yarısında değerlendirme alanında "sınav kültürü" olarak adlandırılan psikometrik modelden "değerlendirme kültürü" ne doğru bir paradigma değişikliği ortaya çıkmıştır (10,74).

Sınav kültürünün sahip olduğu değerlendirme anlayışında, öğrenme ve öğretme ilişkisinin doğal bütünlüğü içinde yer alması gereken değerlendirme, bu bütünlüğün dışına taşınarak ayrı bir bileşen haline gelmiş, hem öğrenen hem de öğreten için ayrı bir anlam ve işlev kazanmıştır (10,80-82). Öğretme ve değerlendirmenin birbirinden ayrı iki öğretim etkinliği olarak ele alındığı "sınav kültüründe", eğiticilerin görevi öğretmekle sınırlı kalırken; soru ve sınavlar başka bir grup eğitici tarafından uzman kişilerin teknik katkı ve destekleriyle hazırlanmış ve sınav sonuçları psikometrik modele göre analiz edilmiştir (13,83). Baird'e göre teknokratik psikometrik yaklaşımlar, öğrenme ve değerlendirmeyi yanlış yönlendirmiştir (67). Eğitim sürecinin gerçek amacı olan öğrenme ve öğrenmenin niteliği göz ardı edilmiş, sınavlardan alınan notlar öğrenmenin kendi gerçekliğinin önüne ve yerine geçerek, amaç ve araç yer değiştirmiştir (84,85). Öğrenmeye dair değerlendirme, belli bir sürenin sonunda yapılan sınavlarda verilen doğru yanıt sayıları, alınan puanlar ve istatistik analizler aracılığıyla açıklanıp yorumlanmış (75); öğrenme bağlamını kaybetmiş, sayılara indirgenmiştir (67). Bireyin performansı ve öğrenmesi hakkında ne öğrendiğimizin tersine, değerlendirme yöntemleri üzerindeki aşırı vurgu değerlendirmenin teknolojik yanını ön plana çıkartmış ve değerlendirme söylemi, teknoloji söyleminin bir parçası olmuştur (23). Oysa değerlendirmenin, öğrencilerin sadece herhangi bir konuda sorulana cevap vermeleri ya da bekleneni yapabilmeleri olarak basit bir biçimde görülemeyeceği, öğrenmeye ve değerlendirilmeye dair tüm deneyim ve algıların değerlendirme sırasında ortaya çıkan sonucu belirlediği vurgulanmıştır (75).

Boud, değerlendirmenin daha rasyonel, etkili ve teknik olarak savunulabilir olmasının, "ölçüm" temelli değerlendirme yaklaşımının temel meselesi olduğuna işaret etmiştir (75). Gipps'e (28) göre, belirleyiciliği yüksek önemde olan, karar verdirici, ölçüm temelli sınavların meşruiyeti, eğitim ve öğrenme üzerinde istenen etkilerinin kanıtlanarak gösterilmesinden değil, eğitim programı, bireyler ve toplum üzerinde yarattıkları algı ve bir kontrol sembolü

olmalarından kaynaklanmıştır. Ölçümden bağımsız olarak düşünülmeyen değerlendirme, psikometrik bakış açısıyla ele alınmış ve geçerlik tartışması ağırlıklı olarak psikometri alanının kavramları üzerinden yürütülmüştür (11,75,86). Psikometrik yaklaşımda güvenirlik, standardizasyon ve nesnellik ile sağlanmıştır (10). Knight, güvenirliği yüksek standardize, nesnel sınavlardan elde edilen notların, yükseköğrenim için ikinci derecede önem taşıyan basit kazanımlar hakkında güvenilir bilgi verebildiğini; bu tür sınavların, karmaşık durumlardaki performansı öngörmeye yönelik gerçek ve özgün bir değerlendirme için uygun olmadığına işaret etmiştir (87).

Yirminci yüzyılın ikinci yarısında sanayi toplumundan bilgi toplumuna geçişle birlikte öğrenme kuramlarında da önemli gelişmeler kaydedilmiş, bilişsel öğrenme kuramı, konstrüktivist öğrenme kuramı ve sosyal konstrüktivist öğrenme kuramları geliştirilmiştir (51,67). Psikometri alanında da klasik test kuramına dayalı ölçüm modellerinin dışında örtük özellikler kuramı (latent trait theory) ve temsiliyet kuramına (representational model of measurement) dayalı ölçüm modelleri kullanılmaya başlanmıştır (9,61,67).

Delandshere, eğitim ve değerlendirme sistemine doğrudan etkisi olan bu gelişmelere karşın, net tanımlanmış "değerlendirme kuramlarının" ortaya çıkmadığını, "ölçme kuramları" ve modelleri çerçevesinde ele alınan değerlendirmenin, daha çok, öğrenenlerin ne bildiklerine karar vermek ve bu kararı verebilmek için gereken verinin nasıl toplanacağı ile sınırlı kaldığını belirtmiştir (23,74). Eğitim alanında öğrenme kuramları ve psikometri alanında ölçme modelleriyle ilişkili önemli gelişmeler olmasına karşın, değerlendirmeye (assessment) ilişkin tanımlanmış kuram ve modeller olmadığı, değerlendirme uygulamalarına dayanak oluşturacak felsefi ve kuramsal tartışmaların yapılmadığı görüşünde olan yazarlar, değerlendirme uygulamalarında karşılaşılan sorunlara egemen değerlendirme anlayışı doğrultusunda çözüm arandığına işaret etmişlerdir (23,34,51,61,66). Baird ve arkadaşlarına göre, psikometrik yaklaşım üzerinden bir öğrenme kuramı ortaya çıkmadığı gibi; psikometri, öğrenme kuramlarının araştırılması ya da geliştirilmesinde temel araştırma tekniği de olmamıştır (67). Psikometrik değerlendirme anlayışı ve nesnel, standardize sınav yaklaşımının temel olarak davranışçı kurama dayandığı farklı yazarlar tarafından sıklıkla belirtilmiştir (9,10,33,34,67,88). Davranışçı bakış açısına göre öğrenme, ardışık, aşamalı ve belli parçalara bölünmüş biçimde organize edilerek eğitici tarafından aktarılan bilginin biriktirilmesi olarak görülmüş; bu anlayış doğrultusunda değerlendirme de, öğrenilenlerin öğrenildiği biçimde tekrar edilmesi ve gözlemlenmesi temelinde, sorulan sorulara verilen doğru cevaplara göre ne kadar bilgi edinildiğinin ölçülmesine dayandırılmıştır (74,82,88). Ölçümün bilimselliği ve güvenirliği için nesnellik ve standardizasyon üzerinde durulmuştur (10,11,28,82).

Mislevy, günümüzde eğitim alanındaki ölçmeyi belirleyen "test teorisini", 20. yüzyıl istatistiklerinin 19. yüzyıl psikolojisine uygulanması olarak tanımlamıştır (14,51). Günümüzde kullanılan sınavların ve eğitim alanındaki psikometrik uygulamaların düşünsel bağlamının 20. yüzyılın ilk yıllarından kaldığı ve geçtiğimiz yüzyılın başlangıcında değerlendirme uygulamalarına öncülük eden psikometrik ilkelerin günümüzde de sürdüğü belirtilmiştir (36,58). Pellegrino, Chudowsky ve Glasser'e göre, önceki, kuramların temel kavram ve modelleri belli amaçlar için hala kullanılabilir olsa da, yeni değerlendirme ihtiyaçlarına cevap verebilmek üzere değiştirilmeleri ya da yeni kuramların geliştirilmeleri gereklidir (14).

Biggs ve Tang, eğitim alanında değerlendirmeye ilişkin kullanılan ölçme modelinin, bilginin sayısallaştırılabileceği, ölçme yoluyla yapılan notlandırmanın evrensel bir anlam taşıdığı,

değerlendirmede psikometrik ve sayısal yaklaşımın bilimsel, kesin ve nesnel olduğu varsayımları üzerine kurulduğuna işaret etmişler ve bu varsayımların doğruluğunu tartışmışlardır (64). Benzer biçimde Gipps (10), değerlendirmenin bir bilim olmadığını ve değerlendirmenin bu biçimde ele alınmasından vazgeçilmesi gerektiğini belirtmiştir.

Psikometrik yaklaşımın, öğrenmenin kapsam ve niteliğini anlamaya ve öğrenmeyi desteklemeye yönelik değerlendirmede uygun olmadığı öne sürüldüğü gibi (11,28,44,51,52), psikoloji alanında kullanılan ölçme modelleri ile eğitim alanındaki değerlendirmeler arasında temel farklılıklara da dikkat çekilmiştir (8,10,11,52,64,66,82,89). Bu farklılıklar şöyle özetlenebilir:

1- Zekâ, kişilik gibi gözlenemeyen sabit örtük bir özelliğin (fixed latent traits) varlığından hareketle, bireylerin bu özelliğe göre nasıl farklılaştığı temelinde psikologlar tarafından geliştirilen ve psikoloji alanında araştırma ve tanı amaçlı olarak kullanılan ölçme modeli, eğitimde ölçmeye konu olan özellik açısından farklıdır. Eğitim alanındaki değerlendirmelerde ölçmenin konusu, örtük değişmeyen kişisel özellikler değil, öğretim ve öğrenme yoluyla kazanılan ve değişmesi beklenen bilgi, beceri ve yeterliklerdir (10,28,64,66,82,90). Psikometrik yaklaşım, test maddelerinin soyut ve örtük bir alanı temsil ettiği varsayımına dayanır; oysa eğitim alanında kullanılan sınavlarda, değerlendirmenin kapsamı eğitim programı çerçevesinde belirli ve somuttur (66).

2- Psikoloji alanındaki testlerde ölçülmek istenen örtük özelliğe ilişkin bir kuram mevcuttur ve ölçülmesi beklenen "yapı" (construct) kuramsaldır, testten alınan puanların toplamının bu kuramsal yapıyı temsil ettiği varsayılır; daha yüksek puan o özelliğe daha fazla sahip olunduğu anlamına gelir (49,66,91). Oysa eğitim alanında başarı testleriyle ölçülen özellik ile ölçüm arasındaki ilişkiyi ve testten alınan puanların temsil ettiği yapıyı açıklayan bir kuram ortaya konmamıştır; sorular üzerinden atanan puanlara göre daha yüksek puan alınmış olunması, örtük bir yapıya sahip oluş derecesini değil, sadece test sorularının doğru yanıtlanmış olduğunu gösterir (66,67,91). Baird ve Black'e göre (66), nesnel, standardize testler ile neyi ölçmeye çalıştığımız çok açık değildir; ölçülen özelliğe ait sağlam bir kuramsal yapı gösterilemiyorsa ya da zayıf bir kuramsal ilişki söz konusuysa ölçümler yoluyla elde edilen sayısal sonuçların yorumlanması tartışmalı hale gelmektedir.

3- Psikoloji alanında test yoluyla elde edilen puanlar, ölçülmek istenen özelliğe sahip oluş derecesini gösteren evrensel (universality) bir anlama sahiptir ve bu açıdan kullanılıp yorumlanmaktadır (10). Benzer biçimde, eğitim alanında psikometrik değerlendirme yaklaşımıyla, öğrenilenler nota dönüştürüldükten sonra toplamalar yapılıp ortalamalar alınmakta ve farklı konu alanları arasında, öğrenciler arasında, farklı zamanlar arasında karşılaştırma yapılmaktadır. Oysa eğitim alanında kullanılan testlerin, psikoloji alanında kullanılan testler gibi, tek bir özelliğin ya da belli özelliklerin ölçümü üzerinden evrensel değerler üretebilmesi tartışmalıdır (10,13,64). Herhangi bir not ya da derecenin eğitimde evrensel genellenebilir bir anlamı olabileceğine itiraz eden değerlendirme yaklaşımına göre, eğitimin amaçları, öğrenme öğretme süreçleri, değerlendirme sonuçları birbirinden ayrı ele alınamaz (13,82).

4- Bireylerin belli bir özelliğe sahip oluşunu birbirlerine göre değerlendiren ve bağıl değerlendirmeyi kullanan ölçme modelleri, bireyler arası karşılaştırmayı ve farklılıkları temel alır (13,54). Eğitim alanındaki değerlendirmelerde ise amaç, bireyler arasındaki farklılıkları ortaya koymak değil, öğrenilmiş olması beklenenler üzerinden "bireyin kendisi" hakkında değerlendirme yapmak, güçlü ve zayıf noktalarını belirleyip eğitim sürecindeki gelişmesinde kendisine yardımcı olmaktır;

bireyin diğer bireylere göre durumu, öğrenme sürecine yönelik değerlendirmelerde değil, seçme sınavlarında dikkate alınabilir (10,28,82).

5- Tek boyutluluk (unidimensionality) varsayımı üzerine dayanan psikometrik ölçüm modelleri için bu varsayımın karşılanması psikoloji ve eğitim alanında farklılık gösterir. Psikoloji alanında, tek bir özellik/yapıyı ölçmeyi hedefleyen testlerde testin tüm maddelerinin sadece bu yapıyı ölçebilecek nitelikte olması beklenir ve istatistiksel olarak, bu özelliğe sahip olmadığı belirlenen maddelerin testten çıkarılmasıyla tek boyutluluk sağlanabilir (66,72). Eğitim alanındaki başarı testlerinde ise sadece tek özelliğin ölçülmesi beklenen bir durum değildir; maddelerin içeriği ve ölçülen özellik açısından, başarı testlerinin tek boyutluluk varsayımını karşılaması genellikle mümkün de değildir (10,66,72).

6- Madde yanıt kuramına dayalı modeller, maddelerin birbirinden bağımsızlığı (item independence) varsayımını gerektirir; bir maddeye verilen yanıtın doğru olma olasılığı, diğer maddelere verilecek yanıtlardan bağımsız olmalıdır (66). Oysa bir başarı testinde, herhangi bir konuya ilişkin, bilginin kullanılması, analiz gibi üst düzey düşünme becerilerini gerektiren bir maddenin, diğer bazı maddelerle bağlantılı olması mümkündür (66).

7- Psikoloji ve eğitim alanında kullanılan testlerde yer alan maddelerin test için seçilme gerekçeleri farklıdır (66). Psikometrik ölçme modellerinde, testin modele uygunluğu için gereken tek boyutluluk, birbirinden bağımsızlık gibi gerekliliklerin yanısıra, maddelerin güçlük ve ayırıcılıkları açısından ölçülmesi hedeflenen özelliğe sahip oluş derecelerine göre bireyleri ayırabilmesi (discrimination) beklenir. Tüm bu özellikler açısından istatistiksel olarak uygun bulunmayan sorular testten çıkarılır (52,66,72,89). Oysa eğitim alanındaki değerlendirmeler eğitim programı ve öğrenme süreci ile ilişkilidir; sınavın kapsamı ve sınavı içeren maddeler istatistiksel özellikleriyle değil, eğitsel özellikleriyle belirlenir (10,64,66).

8- Testin uygulandığı grubun, ölçülen özelliğe sahip oluş derecesi açısından normal dağılım gösterdiği varsayımına dayalı ölçme modellerinin bu varsayımı, yetenek, kişilik, ilgi, zekâ gibi özelliklerden farklı olarak eğitim sürecinin bizzat belirleyici rol oynadığı başarı testleri için tartışmalıdır. Eğitim sürecinde kullanılan başarı testlerinde, testten alınan puanların normal dağılım dışında özellikler sergilemesi sıklıkla görülen ve beklenen bir durumdur, test sonuçlarının dağılımına yönelik bir varsayım genellikle mümkün değildir (10,64,66).

9- Psikometrik yaklaşımla yapılan ölçümler, var olan bir özelliğin fiziksel nesneler gibi bağımsız biçimde ölçülebileceği yönündeki epistemolojik varsayıma dayalıdır; oysa sınavlarla ölçülen özellik sosyal olarak inşa edilen bir özelliktir; sınavların kendisi, ölçülen bu özellik üzerinde biçimlendirici role sahiptir (11,66). Knight, epistemolojik açıdan, öğrenerek kazanılanlar/başarı (achievement) ve yeterliklerin ölçülebilir özellikler olmadığını ve "ölçülemeyenin" ölçülmeye çalışıldığını belirtmiştir (8).

Eğitim alanındaki değerlendirmelerde psikometrik yaklaşımın sorunlu etkisi ve değerlendirme uygulamalarının sadece psikometrik kavramlarla açıklanamayacağı 1980'li yıllardan sonra çok sayıda yazar tarafından önemle vurgulanmış (10,36,64,82,92,93); psikometrik ölçme modellerinin eğitim alanındaki değerlendirmeye uymadığına ilişkin tartışmalarla birlikte değerlendirmeye ilişkin paradigma değişikliği gündeme gelmiştir (10). "Sınav kültürü" olarak adlandırılan, rasyonel paradigmaya dayalı, ölçüm-psikometri odaklı değerlendirme yaklaşımından; yorumlayıcı paradigma çerçevesinde, asıl kullanıcısı öğrenen olan ve öğrenmeyi destekleyen, "edumetrik" değerlendirme olarak da adlandırılan anlayışa dayalı "değerlendirme kültürü"ne geçişten söz edilmiştir (52,81,83,93-95).

Bilgi üretiminin hızla arttığı 21. yüzyılda eğitimin çerçevesi, daha fazla miktarda bilgi edinilmesine yönelik olmaktan çok, öğrenmenin derinliği üzerinden ele alınmaktadır. Günümüzde, bilgiye ulaşma, farklı konu ve bağlamları ilişkilendirme ve karşılaştırma, sorgulama, eleştirel düşünebilme, hipotez geliştirme, kanıtlama, görüş oluşturabilme gibi özelliklerin kazanılması, kazanılan bilgi miktarından daha önemli hale gelmiş; eğitimde, genişliği fazla ama derinliği az olmakla eleştirilen eski anlayış, "daha az, daha yüksek, daha derin" (fewer, higher, deeper) yaklaşımı çerçevesinde değişmeye başlamıştır (96). Bu bağlamda, değerlendirme, niceliksel olarak ne miktarda öğrenildiğine (assessment of learning) ilişkin karar verdirici (summative) kullanımının dışında, öğrenmeye ilişkin niteliksel açıklama getirebilen ve öğrenmeyi destekleyen (assessment for learning) biçimde, geliştirici (formative) işlevleri ile ele alınmaya başlanmıştır (15,44,67,75,89,97). Psikometrik özellikler ile sınırlı teknik rasyonel bir değerlendirme kavramsallaştırmasının ötesinde, değerlendirmenin aynı zamanda sosyal bir süreç olma özelliği ön plana çıkmış; özgün (authentic), durumsal (situational) değerlendirme konuları önem kazanmıştır (9,11,12,28,34,67,77,86). Değerlendirmeye dair yaklaşım değişikliği, değerlendirmeyi "teknik bir iş" olmanın ötesine taşıyarak, değerlendirmenin amacı, işlevi, kullanım biçimleri, etkisi ve yol açtığı sonuçlar geçerlik bağlamında dikkate alınan temel konular arasında yer almıştır (23,28).

5. Testler ve Geçerlik

Yirminci yüzyılın başlangıcı olan 1900 yılı, geçerlik kuramlarının tarihi açısından da başlangıç yılı olarak görülmüştür (98). Psikoloji, endüstri ve eğitim alanında nesnel, standardize testlerin kullanımının yaygınlaşmasıyla beraber, ilk geçerlik tartışmaları ve geçerliğe ilişkin akademik yayınlar, ağırlıklı olarak bu testler bağlamında 20. yy başında ABD'de ortaya çıkmış; "ölçme"nin psikoloji ve eğitim alanında ayrı bir profesyonel alana dönüşmesiyle, geçerlik ve geçerliğin gösterilmesine ilişkin öne sürülen kuram ve kavramlar tüm dünyada yayılıp evrensel nitelik kazanmıştır (32). APA tarafından ilk kez 1954'te "Technical Recommendations for Psychological Tests and Diagnostic Techniques" adıyla yayınlanıp, 1966'da "Standards for Educational and Psychological Tests and Manuals" olarak değişen adıyla güncellenen ve APA, AERA, NCME tarafından 1974, 1985, 1999, 2014 yıllarında beş kez güncellenmiş olan "Standartlar", tüm dünyada geçerlik tartışmalarının yönünü ve içeriğini etkilemiştir (16,32,99-101). Eğitim alanında ölçme kuram ve uygulamaları için "kutsal kitap" sayılan, ilk kez 1951 yılında basılıp 1971, 1989 ve 2006 yıllarında farklı editör ve yazarlar tarafından yayına hazırlanan "Educational Measurement" adlı dört ayrı kitaptan oluşan seri, geçerliği ele alan özel bölümleriyle geçerlik konusu için de temel kaynak olmuştur (16,32). 1950'den günümüze kadar yayınlanan dört farklı geçerlik bölümü, geçerlik anlayışına yön veren dört temel isim olan Cureton (1951), Cronbach (1971), Messick (1989) ve Kane (2006) imzasıyla yayınlanmıştır.

Aynı zamanda bir epistemolojik duruşu da yansıtan geçerlik kuramları, başlangıcından bugüne kadar değişen epistemolojik varsayımlarına göre geçerliğe ilişkin farklı kavramsallaştırmalar ortaya koymuştur (16). Sosyal, kültürel, siyasi gelişmeler, değerlendirme yaklaşımları ve uygulamalarındaki değişimlere paralel biçimde, geçerlik konusunda da, kantitatif ve pozitivist bir anlayıştan, uygulamanın bağlam ve amacına göre farklı türde kanıtların yorumlanmasına dayalı yorumlayıcı anlayışa doğru bir değişim izlenmiştir (16). "Standartlar" ve "Educational Measurement" içinde yer alan geçerlik bölümleri tarihsel sırasıyla ele alındığında bu değişimi izlemek ve anlamak da kolaylaşmaktadır. Newton ve Shaw geçerlik kuramlarının

Tablo 1: Tarihsel süreçte geçerlik anlayışları

Dönem	1900-1952	1954-1974	1975-1999	2000-
Sembol isim Temel kaynak	Cureton 1954 Standartları, Educational Measurement 1971 "Validity" Chp.	Cronbach 1966, 1974 standartları, Educational Measurement 1971 "Test Validation"Chp.	Messick 1985, 1999 Standartları, Educational Measurement 1989 "Validity"Chp.	Kane 2014 Standartları, Educational Measurement 2006 "Test Validation"Chp.
Kuram ve adlandırma	Geçerlik yaklaşımları Ölçüte dayalı "geçerlik yaklaşımı"	Parçalı geçerlik -Trinitarian model Geçerlik "tipleri"	Üniter model (yapı geçerliği) Geçerliğe ilişkin "kanıt tipleri"	Argümana dayalı üniter geçerlik modeli Yorum ve Kullanım Argumanları (IUAs)
Geçerlik kavramsallaştırması	Ölçütle ilişkili korelasyon	Bilimsel kuram ile açıklama ve ilişkilendirme (substantive theory)	Evaluative judgement Bilimsel sorgulama (scientific inquiry)	Bağlama göre ölçmeye ilişkin çıkarımların gerekçelendirilip, kanıtlanması (practical argument)
Tanım	"Test geçerliğine ilişkin temel soru, testin kullanıldığı amaca ne kadar hizmet ettiğidir. Aynı test farklı amaçlar için kullanılabilir ve testin geçerliği bir amaç için yüksek, diğeri için ortalama seviyede, üçüncü bir amaç için düşük olabilir." (Cureton,1951)	"Dar anlamıyla geçerlik, test puanları üzerinden yapılan belli bir çıkarım ya da kestirimin doğruluğunun araştırılması sürecidir… Daha geniş anlamıyla, geçerliğin gösterilmesi, bir test üzerinden yapılan kestirimler, tanımlayıcı ve açıklayıcı bütün yorumların sağlamlığını araştırmaktır" (Cronbach, 1971)	"Test sonuçları ya da diğer tür değerlendirmeler üzerinden yapılan çıkarımların ve kararların uygunluk ve yeterliliğinin kuramsal dayanak ve ampirik kanıtlar tarafından desteklenme derecesine ilişkin yapılan değerlendirmelerin birleşiminden oluşan bir yargıdır" (Messick, 1989)	"Geçerliğin gösterilmesi, yorum ve kullanıma ilişkin argümanların uygunluk ve yeterliliğinin; bu argümanların temelini oluşturan varsayım ve çıkarımların kabul edilebilirliğinin destekleyen ve desteklemeyen kanıtlarla birlikte değerlendirilmesidir. "(Kane, 2013)
Geçerlik-test-yorum ilişkisi	Testin geçerliği	Test puanları üzerinden yapılan yorumların geçerliği	Test puanlarının yorumlanma ve kullanılmasının geçerliği	Test puanlarının yorum ve kullanımına dayanak oluşturan argümanların geçerliği
Geçerliğin gösterilmesi	Mantıksal yaklaşım (logical approach) Ampirik yaklaşım (empirical approach)	1954-1966 - İçerik geçerliği –mantıksal yaklaşım - Eş zamanlı geçerlik –ampirik yaklaşım - Yordama geçerliği – ampirik yaklaşım - Yapı geçerliği 1966-1974 - Yapı geçerliği – mantıksal, ampirik - Ölçüt geçerliği – ampirik, kuramsal - İçerik geçerliği –mantıksal, ampirik	Kanıt tipleri (types of validity evidence): - test kapsamına dayalı kanıtlar (evidence based on test content) - cevaplama süreçlerine dayalı kanıtlar (evidence based on responce processes) - testin iç yapısına dayalı kanıtlar (evidence based on internal structure) - testin diğer değişkenlerle ilişkisine dayalı kanıtlar (evidence based on relations to other variables) - test uygulamasının sonuçlarına dayalı kanıtlar (evidence based on consequences of testing)	- Puanlama: gözlemden gözlemlenmiş puana (from observation to observed score) - Generalization: gözlemlenmiş puandan evrensel puana (from observed score to universal score) - Extrapolation: evrensel puandan hedef alana (from universal score to target domain) - Karar verme: hedef alandan ölçülen özelliğe (from target domain to construct)
Temel kavramlar	Criterion, correlation, prediction, correlation coefficient,	Postulated attribute, hypothetical constructs, nomological network, strong programme, weak programme, theoretical justification	Interpretation/use, ethical considerations, consequences, validity evidence, construct-underepresentation, construct-irrelevant variance, integrated evalative judgement	Interpretive/use arguments, validity arguments, observed attributes, theoretical attributes, inferences, assumptions, claims, warrants, backing
Epistemoloji	Pozitivizm, operasyonalizm	Mantıksal pozitivizm	Konstrüktif realizm	Yorumlamacılık, pragmatizm
Sosyal, Politik, Bilimsel Bağlam	Sanayileşme, verimlilik, eğitimin kitleselleşmesi, test kullanımının artışı, I. Dünya Savaşı, geçerlik kavramının tartışılmaya başlanması	II. Dünya Savaşı, soğuk savaş, eğitim reformları, testlerin standardizasyonu-kontrolü, artan bilimsel araştırmalar	Ayrımcılığa karşı Yurttaş Hakları Hareketi, Medeni Haklar Yasası, pozitivizme yönelik eleştiriler, testlerin yanlış kullanımı, test uygulamalarının olumsuz etki ve sonuçları	Sosyal bilimlerde paradigma değişikliği, farklı değerlendirme yaklaşımları, kuram (validity) ve uygulama (validation) arasındaki kopukluk

gelişimini ve test uygulamalarına temel oluşturan geçerlik anlayışlarını dört döneme ayırmışlardır (32). Farklı yazarlar tarafından da benzer biçimde ele alınan geçerlik anlayışları ve tarihsel süreci, Tablo -1'de özetlenmiştir.

Geçerlik kuramları açısından erken dönem olarak nitelenen 1900-1950 yılları arasındaki geçerlik tartışmaları, ölçüte dayalı geçerlik yaklaşımının ağırlık kazandığı çerçevede yürütülmüş olup, Educational Measurement'in ilk baskısında (1951) "Validity" bölümünü kaleme alan Cureton'un anlayışı, bu döneme egemen olan geçerlik kavramsallaştırmasını da ortaya koymuştur. Bu dönemde geçerlik, ağırlıklı olarak korelasyona dayalı ampirik yaklaşım zemininde tartışılmıştır (17,32,62). Cureton'un geçerlik yaklaşımını "tanımlayıcı ampirisizm" (descriptive empiricisim) olarak niteleyen ve bu yaklaşımın davranışçılığı ve pozitivizmi yansıttığını öne süren Markus ve Borsboom'a göre, gözlemlenebilir davranışlar üzerinden yapılan korelasyon çalışmaları, bu anlayışın hayata geçirilmesine aracılık etmiştir (98). Ölçüte ilişkin kestirim (prediction), bu dönemin temel karakteristiği olarak sayılmıştır (17,32,62,98,102).

Cronbach ve Meehl'in ağırlıklı etkisinden söz edilen 1952-1974 arasındaki dönemde, geçerlik, ayrı "tip" geçerlikler olarak kavramsallaştırılmış, yeni bir kavram olarak "yapı geçerliği" doğmuş ve gelişmiştir (17,32). 1954 yılında yayınlanan Standartlar'da kapsam geçerliği (content validity), yordama geçerliği (predictive validity), eş zamanlı geçerlik (concurrent validity) ve yapı geçerliği (construct validity) olmak üzere dört ayrı geçerlik tipi tanımlanmıştır (99,100). Eş zamanlı geçerlik ve yordama geçerliği, 1966 ve 1974 yıllarında güncellenen Standartlar'da, ölçüt geçerliği başlığı altında birleştirilmiş; kapsam geçerliği, ölçüt geçerliği ve yapı geçerliğinden oluşan "trinitarian model" ya da "kutsal üçleme" (Holy Trinity) ortaya çıkmıştır (32,62). Özellikle yapı geçerliği bağlamında, ölçülmesi hedeflenen "yapı"ya ilişkin güçlü kuramsal açıklama ve ilişkilendirmelerin yapılması gereğinin vurgulandığı bu dönemin geçerlik anlayışında, hem ampirik kanıtlar hem de mantıksal analiz (logical anaysis), geçerliğin gösterilmesi için gerekli görülmüştür (32,62,102-104). Bu dönemde, mantıksal pozitivizmin belirleyici rol oynadığı ve kestirim temelindeki geçerlik anlayışından "açıklama" (explanation) temelinde geçerlik kuramına geçildiği farklı yazarlar tarafından belirtilmiştir (17,32,62,98,103). Markus ve Borsboom, bu dönemin geçerlik yaklaşımını "açıklaycı ampirisizm" (explanatory empiricism) olarak nitelendirmiştir (98).

"Messick yılları" olarak anılan 1974-1999 arası dönemde, geçerliğin tamamının aslında yapı geçerliği olduğu yönündeki kavramsallaştırma egemen olmuş ve "unitarian model" olarak geçerlik tipleri, yapı geçerliği altında birleştirilmiştir (32,102). Birleşik geçerlik kuramında, gözlem yoluyla elde edilen puanlar üzerinden gözlenemeyen özelliklere/yapılara ilişkin yapılan yorumlar ve "yapı" geçerliği" esas alınmış; geçerlik, test puanlarının "yorumu" ile ilişkilendirilmiştir (17). "Test puanlarının yorumu", test puanlarından çeşitli sonuçlara varmaya, bu sonuçlara göre karar vermeye kadar olan aşamalarda yapılan çıkarımların ve varsayımların tümü olarak tanımlanmıştır (104,105). Geçerliğin gösterilmesi (validation), içerik, yanıtlama süreçleri, testin iç yapısı, diğer değişkenlerle ilişkisi ve test uygulamasının yol açtığı sonuçlar olmak üzere beş ayrı başlık temelindeki kavramsal ve ampirik kanıtların ortaya konmasına dayandırılmıştır (32,103). Messick'in yaklaşımında, etik boyutun da geçerliğin bir parçası olarak ele alınması ve testin yol açtığı sonuçların araştırılmasının gereği önemle vurgulanmıştır (17,32). Messick'in geçerlik kuramının, psikoloji ve eğitim alanındaki geçerlik tartışmalarının zeminini mantıksal pozitivizmden (logical positivism) konstrüktivist gerçekçiliğe (constructivist realisim) doğru değiştirdiği ifade edilmiştir

(98,103,106).

2000 sonrasından günümüze kadar olan dönem ise, Kane'nin yorumlamacı anlayış temelinde geliştirdiği geçerlik kuramı ve öne sürdüğü metodolojinin genel kabul gördüğü bir dönem olmuştur (32). Kane'in yaklaşımında geçerlik, test puanlarının kullanımına ve yorumlanmasına temel oluşturan iddia/varsayım/çıkarımların neler olduğunun ortaya konması, bunları destekleyen ve desteklemeyen kanıtların bir arada gösterilmesi (interpretive/use arguments-IUAs) çerçevesinde ele alınmıştır (17,104). Messick'in kavramsallaştırmasında geçerlik, verilmiş bir kurguya göre ortaya konan kanıtların değerlendirilmeleri üzerinden bir yargıya varılması (evaluative judgement) olarak görülürken, Kane'in yaklaşımı, testin geçerliğine ilişkin iddianın nasıl kurgulanıp savunulduğunun gösterilmesini temel almıştır (32). Ortaya konacak kanıtların nitelik ve kapsamının, test sonuçlarının yorumu ve kullanımıyla ilgili iddiaya göre değişebileceğini öne süren Kane'in anlayışı, evrensel standartlar tanımlayan, formal bilimsel "yapı geçerliği" kuramının yerine, geçerliğin gösterilmesini testin amacı ve kullanıldığı bağlam çerçevesinde ele alan dinamik bir yaklaşım ortaya koymuştur (17,32,98,104).

5.1. 1900-1950- Geçerliğin Doğuşu, Geçerlik Yaklaşımları

On dokuzuncu yüzyıl sonu ve 20. yüzyılın ilk yıllarında eğitim, psikoloji ve endüstri alanındaki yaygın test uygulamaları üzerinden "testlerin niteliği" üzerine başlayan tartışmalar, bu üç alandaki uygulamalara ilişkin bazı soruları gündeme getirmiştir. Birincisi, eğitim alanında yazılı sınavların başlamasını takiben, okula girişlerde ve başlanacak eğitim programının belirlenmesinde "standardize testlerin" seçme ve yerleştirme amacıyla sıklıkla kullanılması, bu testlerin sonuçlarının doğruluğu/kesinliğine (accuracy) ilişkin ortaya çıkan sorulardır. İkinci olarak, psikoloji alanında "mental testlerin" hızla artması, korelasyon çalışmaları başta olmak üzere istatistik yöntemlerin gelişmesiyle, hem akademik hem de pratik açıdan bu testlerin bireyler arası mental kapasite farklılıklarını ortaya koyabilme nitelikleri araştırma konusu haline gelmiştir. Üçüncü olarak, iş hayatında "bilimsel yönetim" ilkeleri çerçevesinde çalışma dünyasına uygun seçme testlerine artan talep ve eğitim alanında kullanılan testlerle beraber, zekâ, yetenek, başarı gibi özellikleri ölçme idiasıyla çok sayıda test piyasaya çıkmış ve özellikle ABD'de büyük bir "test endüstrisi" doğmuştur. Bu sektörün kontrol edilip düzenlenmesi ve testlerin alıcısı olan kurum ve kuruluşların testlere yönelik doğru seçim yapıp karar verebilmeleri açısından bu testlerin niteliğinin sorgulanması gereği ortaya çıkmıştır (32,37). Yaygın ve ticari olarak pazarlanan bu testler, geçerlik, güvenirlik tartışmalarının da odağında yer almıştır (44).

İlk geçerlik tartışmaları, özellikle 1910-1920 arası dönemde geniş kabul gören "ölçme hareketi" (measurement movement) bağlamında hızlı bir artışla ortaya çıkan zekâ, yetenek ve başarı testlerinin değerlendirilmesi ve kontrolünün sağlanması girişimleriyle başlamış, bu dönemde ağırlıklı olarak kullanılan korelasyon çalışmaları testlerin niteliğine karar vermede önemli rol oynamıştır (32). Test uygulamalarına ilişkin standardizasyonun ve kontrolün sağlanması girişimleri çerçevesinde APA ve 1928 yılında adı AERA olarak değişen National Association of Directors of Educational Research (NADER) bünyesinde başlayan tartışma ve çalışmalarla, geçerlik konusu, kavramsal olarak ilk kez 1915 yılından sonra gündeme gelmiştir (32,37). NADER Standardizasyon Komitesi tarafından 1921 yılında yayınlanan raporda, "testin neyi ölçtüğünün" ve "ne kadar tutarlı ölçtüğünün" belirlenmesi ölçme ile ilgili en önemli iki sorun olarak ifade edilmiş; ilk sorun "geçerlik", diğeri "güvenirlik" olarak adlandırılmıştır. Aynı raporda, bir testten alınan puanlarla aynı özelliğe ait diğer ölçümler arasındaki

ilişkinin gösterilmesinin, o testin geçerliğinin belirlenmesi sorununu çözeceği belirtilmiştir (32,101). "Bir test ya da sınavın ölçmek istediğini ölçebilme derecesi" olarak 1924 yılında Ruch tarafından yapılan tanım, geçerliğe ilişkin öne sürülen ilk resmi tanımlardan birisi olmuştur (15,32). Bu dönemde geçerlik, "sınava ya da teste" ait bir nitelik olarak kavramsallaştırılmıştır (32). Testin uygulanma koşulları, kullanıldığı bağlam, uygulandığı grup ve testin kullanım amacı gibi faktörlerin yapılan ölçümün niteliği üzerinde etkili olduğu ve geçerliğin ölçüme ait bir özellik olarak ele alınması gerektiği Ruch, Lindquist gibi isimler tarafından 1930'larda tartışılmaya başlanmış (15), ancak bu konudaki netlik 1970'lere kadar tam olarak sağlanamamıştır (32,73).

Uygulama alanları ve kullanım biçimleri farklı olan testler için farklı yaklaşımlar ortaya çıkmasına karşın, özellikle I. Dünya Savaşı sonrası sayıları hızla artan zekâ testleri ve seçme yerleştirme uygulamaları bağlamında tartışılmaya başlanan geçerlik, bu dönemde bir korelasyon meselesi olarak ele alınmıştır (32,62). Bir testin ölçmek istenileni ölçüp ölçmediği sorusunun korelasyon ile yanıtlanabileceği görüşü genel kabul görmüş, testin geçerliği, belli bir ölçütle olan ilişkisi bağlamında korelasyon çalışmaları ile araştırılmıştır (107). Ölçütle test arasındaki korelasyonlar üzerine yapılmış çalışmalarla "başarılı" bir uygulamaya örneklik eden Army Alpha testi, bu yaklaşım üzerinde önemli rol oynamıştır (62). Thorndike, Thurstone gibi araştırmacılar tarafından da "test ve belli bir ölçüt arasındaki ilişkinin derecesi" olarak ifade edilen ölçüte dayalı geçerlik yaklaşımı, ilk geçerlik tartışmalarını şekillendirmiştir (15,32). 1920'lere kadar "korelasyonun derecesi" üzerinden kavramsal ve soyut olarak ele alınan ölçüt konusu, "korelasyon katsayısının" kullanılmasıyla somut ölçümlere dönüşmüş ve geçerliğin gösterilmesi ampirik nitelik kazanmıştır (32). Geçerliğin derecesi, korelasyon katsayısının büyüklüğüne göre değerlendirilmiştir (32,61,62,73). 1920-1950 arası dönem, ölçüt ve test ilişkileri üzerine istatistik modellerin geliştirildiği ve ölçüte dayalı geçerlik yaklaşımının "altın standart" olarak kabul edildiği dönem olmuştur (104). Klasik test teorisinin tüm test uygulamaları için temel alındığı bu dönemde, testlerin güvenirliği de korelasyon çalışmaları zemininde yürütülmüş, Spearman-Brown, Kuder-Richardson 20-21 katsayıları testlerin güvenirliği için temel alınmıştır (54,59).

Bu dönemde korelasyon, bir nedensellik ilişkisi gibi ele alınarak, test ile ölçüt arasındaki korelasyon, geçerliğin kanıtı olarak kabul edilmiştir (61,62). Guilford tarafından "bir test, korelasyon gösterdiği herhangi bir şey için geçerlidir" biçiminde özetlenen bu yaklaşım, kuramsal zeminden yoksun (atheoretical) olduğu belirtilerek, "kör ampirisizm" olarak nitelendirilmektedir (32,62). Klasik test kuramının istatistiksel olarak operasyonalist felsefeye hizmet ettiğini öne süren Borsboom'a göre, araştırmacıların geçerlik için ölçüt geçerliği yaklaşımına odaklanması, klasik test teorisinin merkezinde yer alan "gerçek puan" kavramından kaynaklanmıştır (49). Korelasyon ve nedensellik arasındaki farka vurgu yapan Borsboom, kuramsal bir nedensellik ilişkisi kurulmadan, geçerliğin korelasyon temelinde açıklanmasının, psikolojik ölçme kuramında yapılan en ciddi yanlışlardan birisi olduğunu belirtmiştir (61).

Ölçütün ne olması gerektiği, belirlenen ölçütün özellikleri ve geçerliği, güvenirliği gibi konular 1940'lı yıllara kadar çok fazla tartışılmamıştır (32). Eğitim alanında kullanılan testler için eğiticilerin değerlendirmeleri, diğer testlerden alınan puanlar, toplam not ortalamaları ölçüt olarak ele alınırken; işe alımlarda kullanılan testler için bireylerin üstlendikleri görevlerde gösterdikleri gerçek performansa ilişkin yapılan gözlem ve değerlendirmeler ölçüt olarak kabul edilmiştir (17,32,62). Zekâ testleri için Stanford-Binet testi uzun yıllar boyunca

standart kabul edilmiş ve yeni geliştirilen testler Stanford-Binet ile korelasyonlarına göre valide edilmişlerdir (34). Ayrıca Otis Grup Testi, Army Alpha Test, Terman Grup Testi gibi o dönemde diğerlerinden daha nitelikli olduğu yönünde prestije sahip olan grup testleri diğerleri için ölçüt olarak kullanılmıştır (15).

Ölçütle ilişkili geçerlik yaklaşımına ilişkin çalışmalar, ağırlıklı olarak eğitim ve iş hayatında kurumsallaşmış uygulamalara dönüşen seçme-yerleştirme testleriyle ilgili sorunlara cevap niteliğinde yürütülmüştür (11). Shepard, ölçme alanındaki bu döneme ait yayınların, ölçütün niteliğinden çok, kullanılan ölçütün zamanlamasındaki farklılıklara odaklandığını belirtmiştir (62). Seçme ve yerleştirmede temel alınan özelliklerin testler aracılığıyla ölçümü ile bireylerin sonraki performansı arasındaki ilişki üzerinden testlerin geçerliği sorgulanmıştır (12). Ölçülen özellikle ilişkili kestirim gücü bağlamında ele alınan geçerlik (17,32,62), kestirimlerin doğruluğu olarak tanımlanmıştır (17). Özellikle yetenek testleri için, sonraki performansa dayalı "geleceğe" ait bir kestirim ön plana çıkarken; başarı ve zekâ testlerinden alınan puanların, testin uygulandığı zaman diliminde kullanılan bir ölçütten alınan puanlar ile ilişkisi incelenmiş, eş zamanlı geçerlik çalışması (concurrent validation) yapılmıştır (32,62,102).

Ölçüte dayalı geçerlik kavramsallaştırması, testlerin geçerliğinin değerlendirilmesinde kullanılmasının yanı sıra, Army Alpha testlerinin geliştirilmesi sırasında kullanılan "ölçüt" kavramıyla, test geliştirmeye ilişkin anlayış da değişmiştir (32). 1920'lerin ortasından sonra sadece test puanlarının değil, madde puanlarının da ölçütle korelasyonu araştırılmış, maddeler istatistiksel özelliklerine göre bireyler arasındaki farkı ortaya koyabilme nitelikleri üzerinden değerlendirilerek, testte yer alıp almamalarına karar verilmeye başlanmıştır (32).

Eğitim ve iş alanında kullanılan testlerle ilgili olarak gözlenebilen gerçek bir performans olması ve okul ya da işte beklenenlere ilişkin bir kapsam ve örneklemin tanımlanabilir oluşu, testte yer alan maddelerin örneklemi uygun biçimde temsiliyetinin dikkate alınmasını mümkün kılmıştır. Bu alanlarda testlerin değerlendirilmesinde ve testlerin geliştirilme süreçlerinde mantıksal (logical) yaklaşımla, içeriğe ilişkin uzman görüşü, test içeriğinin yürütülen program ya da kaynak kitaplarla uyumu ve ulusal kurulların kapsama ilişkin önerileri dikkate alınmaya devam etmiştir (32,104). Ancak, psikoloji alanında özellikle zekâ, kişilik gibi sınırlı bir kapsam çerçevesinde tanımlanamayan özellikler için, testte yer alan maddelerin örneklemi uygun biçimde temsiliyeti göz önünde bulundurulamayacağından, maddelere ilişkin uzman görüşlerinin yanı sıra, her maddenin aynı özelliği ölçebilmesi ve ayırıcılığı gibi nitelikler ön plana çıkmıştır; testte yer alan maddelerin seçimi teknik ve ampirik bir konu olarak ele alınmıştır (32). Test geliştirmeye ilişkin bu teknik yaklaşım, kısa sürede eğitim ve iş alanlarında da etkisini göstermiştir. Eğitim alanına Lindquist'in öncülüğünde (1936) giren bu yaklaşımla, başarı testlerinde yer alacak maddelerin seçimi için maddelerin testin ilişkili olduğu alanı ne kadar temsil ettiklerinden çok, madde analizleri, genel başarı ölçütüyle madde korelasyonu (yüksek ve düşük başarıya sahip öğrencileri ayırabilmesi), temel alınmaya başlanmıştır (32). Aynı test içinde yer alan maddelerin, testte yer alan diğer bir madde için "ölçüt" sayılmasını, diğer bir deyişle bir testin geçerliğinin testin kendisiyle gösterilmesini, test geliştirme ve geçerlik açısından eleştiren, bu uygulamanın testin güvenirliği ile ilişkili olduğu ve böylece testin geçerliği ile güvenirliği arasındaki ayırımın karıştığı yönünde farklı görüşler öne sürülmüştür (32).

1940'lı yıllara kadar bugünkü anlamda farklı "geçerlik tipleri" değil, farklı "geçerlik yaklaşımları" söz konusu olmuştur (32). Mantıksal (logical) yaklaşım ve ampirik

yaklaşım olarak adlandırılan temelde iki farklı yaklaşım, birinin diğerine göre daha öncelikli görülmesine karşın, ilk yıllarda birbirinin alternatifi olarak değil, birbirini tamamlayan bir anlayışla ele alınmıştır. Geçerlik, ampirik yaklaşım tarafından, istatistik yöntemler, korelasyon çerçevesinde incelenirken; mantıksal yaklaşım, özellikle başarı testleri için eğitim hedeflerinin test kapsamında temsiliyetini göz önünde bulundurmuştur. Testte yer alan maddelerin konu alanını temsiliyetine gösterilen dikkat ile ayırıcılık temelinde madde istatistiklerine verilen önem arasında 1930-1940 yıllarında ayrışma ortaya çıkmış; önemli eğitim hedeflerinin ihmal edildiği, öğrenme ve öğretmeye olumsuz etkileri olduğu ve geçerliğin uygun ölçütlerle gösterilmediği gerekçeleriyle ampirik yaklaşım eleştirilmiştir (32). 1950'li yıllara gelindiğinde mantıksal yaklaşım ile ampirik yaklaşım arasındaki gerilim belirginleşmiş, Thorndike, Rulon gibi yazarlar mantıksal yaklaşımın önemini vurgularken, Guilford, Cureton gibi yazarlar ampirik yaklaşımı ön plana çıkarmıştır (32,62,102). Bu ayrışma, kapsam geçerliği, ölçüt geçerliği gibi farklı "geçerlik tiplerinin" ortaya çıktığı dönemin başlangıcını oluşturmuştur.

5.2. 1950- 1975 Geçerlik Bölünüyor- Farklı Geçerlik Tipleri

Yirminci yüzyılın başından 1940'lı yılların sonuna kadar geçen dönem, geçerliğin kavramsal olarak tartışılmaya ve tanımlanmaya başlandığı, tam bir görüş birliğinin sağlanamadığı kaotik bir dönem olmuştur; farklı görüşler üzerinden uzlaşmanın sağlanması, test uygulamalarına ilişkin kontrol ve standardizasyon konusunda kurumsal çabaların başlaması, 20. yüzyılın ikinci yarısıyla birlikte gerçekleşmiştir (73,101). Bu çabalar, önceki yarım yüzyıllık test uygulamaları üzerinden ortaya çıkan yasal, psikometrik ve pratiğe ilişkin sorunlar bağlamında yürütülmüştür (73).

1940'lı yıllarda, geçerlik yaklaşımları arasındaki farklılıkların, testlerin geliştirilmesi ve kullanımına ilişkin farklı uygulamaların artmasıyla birlikte, 1950-1954 yılları arası Cronbach'ın başkanlığını yürüttüğü APA'ya bağlı "Test Standartları Komitesi", geçerlik tartışmalarında uzlaşmanın sağlanması ve testlerin yayınlanmasından önce sahip olmaları gereken nitelikleri belirlemek üzere bir çalışma başlatmıştır (32). 1952'de taslak öneri olarak sonuçlanan bu çalışmada "predictive", "status", "content" ve "congruent" olmak üzere geçerliğe ilişkin dört kategori tanımlanmıştır (101,108). Bu çalışma sonucu ortaya çıkan taslak öneriler üzerinden AERA ve National Council on Measurements Used in Education (NCMUE) katkılarıyla geliştirilip düzenlenerek kabul edilen ortak görüş, ilk kez 1954 yılında "Technical Recommendations for Psychological Tests and Diagnostic Techniques" adıyla APA tarafından yayınlanmıştır (32,99-101).

"Technical Recommendations" olarak ilk kez 1954 yılında yayınlandığında geçerlik, testlerin farklı amaçlarına göre ele alınmış ve farklı tipteki geçerlik incelemeleri ile ilişkilendirilmiştir (16,62,103). Yayınlanan bu dökümanda, taslak önerilerden farklı olarak, kapsam geçerliği (content validity), yordama geçerliği (predictive validity), eş zamanlı geçerlik (concurrent validity), yapı geçerliği (construct validity) olmak üzere dört geçerlik tipi tanımlanmıştır (32,62,101,104,108). Kapsam geçerliği, önceki mantıksal geçerlik yaklaşımı çerçevesinde; yordama geçerliği ve eş zamanlı geçerlik ise ampirik geçerlik yaklaşımı çerçevesinde ele alınmıştır (32). Bu dönemin geçerlik kavramsallaştırmasında önceki döneme göre iki temel fark ortaya çıkmıştır. Birincisi, "yapı geçerliği" kavramının öne sürülmüş olması; ikincisi de, "yapı geçerliği" tartışmalarıyla birlikte, ilk dönemde ölçüt geçerliği temelinde egemen olan operasyonalizmden, yapı geçerliği temelinde mantıksal pozitivizme doğru bir anlayış değişikliğinin görülmesidir (32,61).

Standartların yayınlanan ilk versiyonunda

geçerlik, parçalı bir anlayışla ele alınarak, eğitimde kullanılan başarı testleri için kapsam geçerliği; yönetsel süreçlerde seçme-yerleştirme amacıyla kullanılacak yetenek testleri için yordama geçerliği; psikoloji alanında tanı amaçlı kullanılacak testlerde eş zamanlı geçerliğin incelenmesi ve belli bir ölçüt ya da ölçüt sayılabilecek başka bir değişken bulunmadığında "yapı geçerliğinin"dikkate alınması önerilmiştir (16,32,104).

5.2.1. Kapsam geçerliği

Testin temsil etme iddiasında bulunduğu durumlara ait evrende bireyin nasıl bir performans göstereceğini belirlemek amacıyla hazırlanan testler, kapsam geçerliği ile ilişkilendirilmiştir (16,32). Kapsam geçerliği, test sonuçları ile varılacak kararları ilgilendiren konu alanı ya da durumların test tarafından ne ölçüde temsil edildiği/örneklendiği olarak tanımlanmıştır (16,103). Üzerinde ustalaşmak üzere çaba gösterilen öğrenme hedeflerini temsil ettiği varsayılan testin, bu hedef davranışları ölçmek üzere hazırlanmasının ve testte yer alan maddelerin temsiliyetini gösteren belirtke tablosunun (test specification) oluşturulmasının gerektiği belirtilmiştir (32). Genel bir yaklaşımla, başarı testlerine ilişkin değerlendirme, kapsam geçerliği ile ilişkilendirilmiş; örneklemin uygunluğu için niteliksel kanıtların gösterilmesinin zor olduğu, bunun yerine kullanılan madde sayılarının uygunluğunun uzmanlar tarafından puanlanabileceğinden söz edilmiştir (32,104). Uzman görüşlerinin yanında, kapsam geçerliği açısından iç tutarlılık katsayısı ve test için verilen sürenin dikkate alınması gerektiği de bildirilmiştir (32).

Shepard'a göre kapsam geçerliğine ilişkin gerekçeler, davranışçılık ve her kavramın gözlemlenip sayısallaştırılacağını öne süren operasyonalizmin etkisinde kalmıştır (62). Tyler'ın da aralarında bulunduğu davranışçı geleneğe bağlı bazı "ampirisizmin körlüğüne" karşı, öncelikle konu alanı ya da "yapı"nın kavramsal analizinin yapılması gerektiğini, sonrasında hedeflenen alanı temsil edecek maddelerin hazırlanması gerektiğini savunmuşlardır (62). Kapsam geçerliğinin değerlendirilmesi için de, test geliştirme süreçlerindeki adımların izlenmesinden söz edilmiştir. Test kapsamının uygunluğunun, eğitim alanında konu alanı uzmanları tarafından değerlendirilmesi önerilirken, iş alanındaki seçmelerde ise endüstriyel psikologların ve iş kolunda çalışan uzmanların görüşlerinin alınmasının uygun olacağı belirtilmiştir (62,73).

5.2.2. Yordama geçerliği

Bir bireyin gelecekteki durumunu tahmin etme ya da bir değişkene göre durumunu kestirme amacıyla hazırlanan testler, ölçüt geçerliği ile ilişkilendirilerek, eş zamanlı geçerlik ve yordama geçerliği olarak iki ayrı başlık altında ele alınmıştır. Test sonuçları ile geleceğe ait bir ölçüt arasındaki ilişkinin gösterilmesi yordama geçerliği olarak ifade edilmiştir (103). İş başvurusunda bulunan bireyin henüz ustalaşmadığı işe ait becerilere ilişkin potansiyelini yordamak üzere hazırlanmış testler ve yetenek testleri bu kategoriye örnek olarak verilmiştir (32). Kapsam geçerliği gibi doğrudan hedef davranışın ölçümüne yönelik değerlendirmenin yapılmadığı yordama geçerliğinde, ölçüte ilişkin davranışı dolaylı olarak işaret eden test davranışının ölçümü ele alınmıştır. Gelecek bir zaman içinde kullanılacak kritere ait ölçümler ile test puanları arasındaki korelasyon, yordama geçerliğinin kanıtı sayılmış; geçerlik katsayısının yanı sıra, puan aralıklarına göre yapılacak yordamanın içerebileceği hata payının göstergesi olarak, olasılık hatasının bildirilmesi gerektiği belirtilmiştir (32).

5.2.3. Eş zamanlı geçerlik

Test davranışlarının dolaylı olarak işaret ettiği belirli özellikler evreni bağlamında, test ile

aynı zaman dilimi içerisinde bireyin beklenen davranışlarının nasıl olabileceğinin belirlenmesi eş zamanlı geçerlik ile ilişikilendirilmiştir (32). Eş zamanlı geçerliğin gösterilmesi için testin uygulandığı zaman içinde kullanılmış bir ölçüte ait puanlar ile test puanları arasındaki korelasyonun araştırılması gerektiği ifade edilmiştir (62). Eş zamanlı geçerlik kategorisinin tahmin/yordamadan ziyade, ayırım yapmakla ilgili olduğu vurgulanarak, tanı ve sınıflama testlerinin geçerliği bu kategoride ele alınmıştır (32). Klinikte depresyon tanısı koymak üzere geliştirilmiş yeni bir ölçeğin önceden kullanılan geçerli ve güvenilir diğer bir ölçek ile birlikte uygulanması bu kategoriye örnek olarak verilmiştir (32). Test sonuçlarının daha kesin sonuç veren eş zamanlı başka bir ölçütle korelasyonunun gösterilmesi eş zamanlı geçerliğin kanıtı sayılmış, ancak bu eş zamanlı geçerliğin gösterilmesinin yordama geçerliği için bir kanıt oluşturmayacağı belirtilmiştir (32).

5.2.4. Yapı geçerliği

Doğrudan gözlenemeyen ve test performansının atfedildiği kavram ya da yapılar aracılığıyla saptanan herhangi bir hipotetik özellik (trait) ya da yapıya (construct), bireylerin ne derecede sahip olduğu konusunda kestirimde bulunmayı amaçlayan testler yapı geçerliği ile ilişkilendirilmiştir. Yapı geçerliğinin testin ölçtüğü niteliklerin, testin dayandığı kuram temelinde, hem mantıksal hem de ampirik olarak farklı çalışmalarla gösterileceği belirtilmiştir (103).

Yapı geçerliğinin ortaya çıkışı diğer üç kategoriye göre farklılık göstermiştir (32). Yapı geçerliği, "Test Standartları Komitesi" tarafından ilk olarak, kişilik testlerinin önceki dönemde uygun olmayan biçimde kuramsallaştırıldığına dair görüşlere dayalı olarak, bu testlerin yorumlanması ile ilgili yaşanan sorunları çözebilmek bağlamında tartışılmaya başlanmıştır (32). Kişilik testleri gibi, teste ilişkin bir ölçüt üzerinden altın standardın olmadığı durumlarda, bu tür testlerin farklı biçimde değerlendirilmesi ihtiyacı doğmuştur (32). Başlangıçta, uyum geçerliği olarak (congruent validity) olarak test sonuçları üzerinden yapılan yorumların uygunluğunun sorgulanmasıyla ele alınan geçerlik kategorisi, sonradan yapı geçerliği olarak adlandırılmıştır. Kişilik testleri bağlamında, ilk kez Cronbach ve Meehl tarafından öne sürülen "yapı geçerliği" kavramı, ilk dönemde, test puanlarının ne anlama geldiğini operasyonel olarak tanımlayacak yeterli ya da uygun belli bir konu alanı ya da ölçüt sayılabilecek başka bir değişken olmadığında, dikkate alınması gereken, "dolaylı" (indirect) bir geçerlik kategorisi olarak görülmüştür (16,32,62,103).

Anket biçiminde uygulanan kişilik testleriyle ilgili olarak operayonalist yaklaşımı eleştiren Meehl ile Cronbach'ın birlikte yürüttükleri çalışmalar, "yapı geçerliği"nin ilk dönem sahip olduğu anlam ve rolü başka bir boyuta taşımıştır (32). Felsefe alanında uzman olan ve aynı zamanda kişilik özelliklerinin ölçümleri ile ilgili geniş deneyime sahip Meehl'in yapı geçerliğinin felsefi altyapısına katkısı ve Cronbach'ın eğitim ve psikoloji alanında testlere ilişkin bilgi ve birikimi zemininde "yapı geçerliği" kavramsallaştırılmış; Cronbach ve Meehl'in bu çalışmaları, sonraki yıllarda geçerlik konusunda temel rol oynayacak kavram ve yaklaşımlar için belirleyici olmuştur (16,32).

Cronbach ve Meehl'in "yapı geçerliği" yaklaşımı, geçerlik anlayışını üç temel noktada değiştirmiştir. Birincisi, ölçülmek istenen yapıya ilişkin ölçüm yoluyla çıkarımda bulunulabilmesi için bu çıkarımların kuramsal olarak açıklanan ilişkilere (nomologic network) dayanması yapı geçerliğinin esasını oluşturmuş; yapı geçerliğinin gösterilmesinde hem mantıksal hem de ampirik yaklaşımın kullanımı gerekli görülmüştür. Bunun için üç temel metodolojik ilke belirtilmiştir. Öncelikle, ölçülmesi hedeflenen yapı ile test durumundaki davranışı açıklayan bir kuramın geliştirilmesi;

ikinci olarak, bu kurama dayalı olarak yapılan çıkarımların belirlenmesi ve üçüncü olarak, veri toplanması ve bu çıkarımların doğruluğunun test edilmesinin gerekliliği vurgulanmıştır (16,17,32).

"Ölçüte dayalı geçerlik"te, geçerlik, önceden var olan bir ölçüte göre değerlendirilmiş ve kullanılan ölçüte ilişkin detaylı bir inceleme yapılmadan, görünüm olarak ölçüt ile testin birbirine benzerliği ve ölçüt ile test sonuçları arasındaki korelasyon dayanak alınmıştır; Cronbach ve Meehl'in "yapı geçerliği" yaklaşımında ise ölçüm sonuçlarının yorumlanmasında, yapılan çıkarımlar ile ölçülen "yapı" arasındaki ilişkinin detaylı incelenmesi gereği özellikle vurgulanmıştır (17). Ölçüte dayalı geçerlik, varolan bir değişkenin ölçümüne bağlı olarak, geçerliği, teste ait bir özellik olarak ele almıştır. Geçerliğin teste ait bir özellik olduğu anlayışı, test puanlarıyla "yapı" arasındaki ilişkilendirme ve çıkarımların kuramsal olarak gösterilmesi ve kanıtlanması bağlamında değişmiştir. Geçerlik anlayışında meydana gelen bu ikinci temel değişim ile beraber geçerlik, ölçüm puanı üzerinden yapılan yorum/çıkarımlara ilişkin bir özellik olarak değerlendirilmeye başlanmıştır (17).

Geçerlik anlayışını etkileyen üçüncü önemli değişim, Cronbach ve Meehl tarafından çok açık biçimde belirtilmemiş olsa da, sonraki yıllarda yapı geçerliği üzerinden gündeme gelen "test puanlarının farklı kullanım ve yorumları"nın tartışılmaya başlanmasıyla ortaya çıkmıştır (17,32).

5.3. Geçerliğin "Kutsal Üçlemesi" – Trinitarian Dönem

1966 yılında "Standards for Educational and Psychological Tests and Manuals" olarak adı değiştirilen "Standartlar"da önceki dönemde "yordama geçerliği" ve "eş zamanlı geçerlik" olarak ele alınan geçerlik tipleri, "ölçütle ilişkili geçerlikler" başlığı altında birleştirilmiş ve "kapsamla ilişkili geçerlik", "ölçütle ilişkili geçerlikler", "yapıyla ilişkili geçerlik" başlıklarıyla üç geçerlik tipi tanımlanmıştır (32,103,108). Guion tarafından "kutsal üçleme" olarak nitelendirilen ve "trinitarian dönemi" tarif eden üç "tip" geçerlik, 1970'lerin sonlarına kadar geçerlik anlayışını şekillendirmiştir (32,62,104).

Trinitarian modelin dayandığı geçerlik kuramıyla, geçerlik tiplerinin birbiri yerine geçemeyeceği, geçerliğin gösterilmesinde her bir geçerlik tipinin ele alınması gerekliliği vurgulanmış olmakla beraber, birbirinden bağımsız, parçalı geçerlik anlayışına dayalı olarak 1950-1970 yılları arasında, seçme testleri için ölçüt, başarı testleri için kapsam, klinik psikoloji alanında ise yapıya ilişkin geçerliğin gösterilmesi uygulaması sürmüştür (16,32). Bu dönemde, ölçütle ilişkili geçerlik için korelasyonel çalışmalar; kapsamla ilişkili geçerlik için uzman görüşü; ve yapı ile ilişkili geçerlik için kuramsal tanımlamalar ile ampirik bulguların tutarlılığı ve uyumu dikkate alınmıştır (103).

1970'lerde Cronbach, Messick, Guion gibi geçerlik kuramcıları, geçerlik tiplerinin farklı uygulamalar için farklı biçimlerde kullanıldığına dikkat çekmiş ve geçerliğin gösterilmesine ilişkin genel bir çerçeve olmadığı için kolaycı bir yaklaşımla, verilerin uygunluğuna göre geçerlik kanıtlarının gösterilmesini eleştirmişlerdir (17,32,104). "Toolkit" olarak adlandırılan bu yaklaşımla, yeteneğe ilişkin "yapı"yı ölçen yerleştirme testleri için "yapı geçerliğinin" gösterilmesi yerine, ölçütle ilişkili ya da kapsamla ilişkili geçerlik kanıtları kullanılmıştır (17,32,104). Kapsama ilişkin analizlerin, testin geliştirilmesi sırasında ya da test uygulamasından hemen sonra yapılması nedeniyle, doğrulamada taraf tutmaya (confirmation bias) açık oldukları da bu yıllarda öne sürülmüştür (104).

Geçerliğin gösterilmesinde "toolkit" uygulamasının başarı testleri için ortaya çıkışını, 1960-70 arası bağıl değerlendirme (norm-

referenced) yerine, ölçüte dayalı (criterion-referenced) değerlendirmenin ve "tam öğrenme" yaklaşımının kullanılmaya başlanmasıyla ilişkilendiren Newton ve Shaw, tüm öğrencilerin eğitim içeriğiyle ilgili ustalık kazanmasının beklendiği bir durumda, öğrenciler arasında ayırıcılığı gösterebilecek korelasyona dayalı ölçüt geçerliğinin anlamlı bulunmadığını belirtmişlerdir (32). İşe alımlarda kullanılan testlerde "toolkit" uygulaması ise 1964 yılında çıkarılan "Medeni Haklar Yasası" ve bu gelişmeyle paralel olarak yargıya taşınan davalar ile ilişkilendirilmiştir (17,32,104). ABD'de eğitim alanı, çalışma hayatı, sosyal ve siyasi yaşamda sürdürülen ayrımcı politikalara karşı 1950'lilerin ortalarında yükselen, 1960'lı yıllar boyunca devam eden Yurttaş Hakları Hareketi, 1964 yılında, okullarda, işe alımlarda ve sosyal yaşamda ırkçı ayrımcılığı yasaklayan bir yasal düzenleme olan "Medeni Haklar Yasası"nın (Civil Rights Act) çıkarılmasına yol açmıştır (32,33,104). Bu gelişmeler doğrultusunda, eğitim ve çalışma hayatında siyahlara ve azınlık gruplarına yönelik ayrımcılıkla ilgili yürütülen tartışmaların önemli konularından birisi test uygulamaları olmuştur (32-34,104). Bu yasaya dayanarak özellikle işe alımlarda kullanılmış olan testlerle ilgili önemli davalar açılmıştır; testlerin savunulabilirliğinin kamuoyunda da sorgulandığı bu dönemde, "Association of Black Psychologists" siyahlara uygulanmış bütün testlerin iptalini talep etmiştir (32). Açılan davalarda, yordamaya dayalı kullanılan bu testlerin kapsamlarının işin gerektirdiği özellikler ile ilişkili olmadığı sıkça tartışılmış ve seçme testlerinde geleneksel olarak yordamanın temel alınması yerine testin kapsamı önem kazanmaya başlamıştır (32).

1954 ve 1966 Standartlarında "geçerlik tipleri" olarak adlandırılan başlık, 1974 Standartlarında değiştirilerek "aspects" nitelemesi kullanılmıştır. Geçerliğin gösterilmesinde farklı seçeneklerin olduğu gibi bir algıyı engellemek amacıyla yapılan bu değişimle, kapsam, ölçüt ve yapıyla ilişkili yanlar olmak üzere, geçerliğin üç farklı boyutu vurgulanmıştır (32,62,103,108). Bu farklı yanların hem mantıksal hem de operasyonel olarak birbiri ile ilişkili olduğu ve geçerliğin gösterilmesinin bunların hepsine ilişkin bilgiyi içermesi gerektiği 1974 Standartları tarafından belirtilmiştir (32).

1970'li yıllardaki geçerlik tartışmalarının önemli sonuçlarından birisi de, test ya da bir değerlendirmede elde edilen puanların kendisine dayalı bir geçerlik anlayışı yerine, belli işlemler sonucu elde edilen puanların yorumlanması, buradan hareketle yapılacak çıkarımlara dayalı bir geçerlik anlayışının kabul edilmesi olmuştur (15). Bu tartışmaların sonucunda, Cronbach tarafından önceki yıllarda öne sürüldüğü gibi, geçerliğin ölçme işlemleriyle elde edilen verilerin yorumlanmasıyla ilişkili bir "süreç" olduğu, testin kendisine ait bir "özellik" olmadığı görüşü kabul edilmiştir (13). 1974 Standartları'na göre geçerlik, "test puanları ya da diğer tipteki değerlendirmelerden hareketle yapılan çıkarımların uygunluğu" olarak tanımlanmıştır (15,62).

Geçerlik kavramına ilişkin erken dönem tartışmalarda, testin ölçmek istediğini ölçebilmesi olarak kavramsallaştırılan geçerlik anlayışı, test puanlarının anlamlandırılması (score meaning) noktasında, yapılan "ölçümün" geçerliği ile yapılan "yordamanın" geçerliği gibi iki farklı noktayı ortaya çıkarmıştır (32). Cronbach ve Loevinger, "yapı geçerliğinin" hem ölçmenin niteliği (descriptive), hem de yordamaya (decision-making) ilişkin değerlendirmenin temeli olduğunu savunmuşlardır (32). Bu görüş üzerinde yürütülen tartışmalarda, yordamanın savunulabilmesinin öncelikle ölçümün savunulabilir olmasını gerektirdiği vurgulanmış; yapı geçerliğinin gösterilmesi, test puanlarının anlamlandırılmasına yönelik "bilimsel araştırma" niteliği kazanmaya başlamıştır (32). 1957'de Loevinger, ölçüt geçerliği ve kapsam geçerliğinin kullanıma özel olması nedeniyle,

bilimsel açıdan, yapı geçerliğinin geçerliğin kendisi olarak ele alınabileceğini önermiş; bu öneri ve yapı geçerliğinin egemen model olduğu "üniter geçerlik görüşü", 1970'lerde geçerlik kuramcıları tarafından genel kabul görmeye başlamıştır (32,104). Bu görüşü savunanlar, kapsam ve ölçüt geçerliğine ilişkin kanıtların ihmal edilemeyeceğini, fakat daha sağlam bir geçerlik çalışması için tüm kanıtların bir arada ele alınmasının doğru olacağını öne sürmüşlerdir (17,103). Bu görüş etrafındaki tartışmalar, geçerliği "üniter" model çerçevesinde ele alan sonraki geçerlik kuramının temelini oluşturmuştur.

5.4. Messick Yılları – Üniter Model

Messick üç ayrı geçerlik olarak parçalanmış geçerlik yaklaşımını ve her bir tip geçerliğin ayrı ayrı değerlendirilmesini eleştirmiş; farklı tipteki geçerlikler için elde edilen çeşitli kanıtların birbirinin alternatifi değil, destekleyicisi olabileceğini ve her türlü geçerlik kanıtının yapı geçerliği kapsamına girdiğini öne sürmüştür (109). Messick'in bu görüşleri ortaya koyması, 1960'lı yıllarla beraber eğitim, psikoloji ve işe alımlarda uygulanan testlere yönelik eleştirilerin yükselmesiyle paralellik göstermiştir (32). Ayrı geçerlik başlıkları üzerinden ele alınan geçerlik anlayışına bağlı olarak uygulamada karşılaşılan sorunlarla ilişkili tartışmaların yürütüldüğü 1970'li yıllardan başlamak üzere Messick, eğitim alanı ya da işe alımlarda kullanılan testler için sadece kapsam geçerliği ya da ölçüt geçerliğinin gösterilmesinin yeterli olamayacağı görüşünü savunmuştur (110,111).

Test içeriği ile ölçülen yapı arasında net bir ayrım yapılamayacağını ve kapsama dayalı olarak yapılan çıkarımlar ile yapıya dayalı olarak yapılan çıkarımların birbirinden ayrılamayacağını ifade eden Messick, kapsam geçerliğinin gösterilmesine karşın, ölçülen yapıdan bağımsız olarak test puanlarında varyans ortaya çıkabileceğine dikkat çekmiştir (109). Ölçülen yapının yetersiz temsiliyeti (construct-underrepresentation) ve ölçülen yapıyla ilgisi olmayan varyans (construct-irrelevant variance) olmak üzere geçerlikle ilgili iki tür tehdit tanımlayan Messick, ilk tehdidi testin gerçek yaşama uygunluğu (authenticity), diğerini ise testin ölçülmesi hedeflenen özelliği doğrudan ölçebilme özelliği (directedness) ile ilişkilendirmiştir (109-111).

Messick, kapsam geçerliği gösterilmiş olsa bile, ölçülen yapıdan bağımsız olarak, testte yer alan soruların yanıtlanabileceği (construct irrelevant easiness) ya da yanıtlanamayacağı (construct irrelevant difficulty) biçimde, test tekniği ya da öğrenci grubuna bağlı olarak değişik nedenlerle (soruların ifade edilişi, test formatı, motivasyon, stress, vb) teste verilecek cevapların farklılaşıp test puanlarının hatalı yorumlanmasına yol açabileceğini belirtmiştir (109-111).

Özellikle işe alımlarda kullanılan testler için temel alınan ölçüt geçerliğine ilişkin olarak Messick, ölçülmesi hedeflenen özellik ile bu özelliği temsil ettiği düşünülen ölçüt arasındaki korelasyonun ampirik olarak gösterilmesinin yordama için yeterli olamayacağını, ampirik kanıtların dışında bu ilişkinin mantıksal ve kuramsal olarak sağlam bir açıklamasının yapılması gerektiğini öne sürmüş, bu noktada, ölçüt geçerliğinin yapı geçerliğinden bağımsız ele alınmayacağına işaret etmiştir (109-111).

1970'li yılların sonlarına doğru, aralarında Cronbach, Guion gibi isimlerin de bulunduğu bir grup geçerlik kuramcısı ve araştırmacı, yapı geçerliğinin önemli rolünün yanısıra test kullanımının yol açtığı sosyal sonuçların önemine dikkat çekmiş (93) ve test uygulamalarının eğitim üzerindeki etkileri ve etik konular üzerinde durmuşlardır (32). Bu tartışmalar bağlamında Messick, konuyu iki soru üzerinden ele almış, "test, ölçmeyi hedeflediği özelliğin ölçümü için iyi niteliğe sahip midir?" olarak belirlediği ilk sorunun, "bilimsel" nitelikli olduğunu ve testin psikometrik özellikleri, özellikle de yapı

geçerliğinin gösterilmesi ile cevaplanabileceğini belirtmiştir (32,110). "Test, amacı doğrultusunda kullanılmalı mıdır?" olarak ifade ettiği diğer sorunun ise "etikle" ilgili olduğuna işaret eden Messick, bu sorunun, değerler açısından testin yol açacağı potansiyel sonuçların değerlendirilmesiyle yanıtlanabileceğini öne sürmüştür (32). Bu iki konu, Messick tarafından birbirinden ayrı olarak ele alınmıştır. Test uygulamasının yol açtığı sonuçları geçerlik kuramının önemli bir parçası olarak gören Messick, test puanlarının yorumlanma ve anlamlandırılması ile testin kullanımını içeren iki yönlü geçerlik kavramsallaştırmasını, 1980 yılında ortaya koyduğu "progressive matrix" ile ifade etmiştir (109-111). (Şekil -1)

Şekil 1: Geçerliğin farklı boyutları (Facets of Validity) (Messick, 1987)

	Test interpretation	Test use
Evidential basis	1 Construct validity (CV)	3 CV+relevance/utility (R/U)
Consequential basis	2 CV+Value implications (VI)	4 CV+VI+R/U Social consequences

Messick, "progressive matrix"te yer alan dört ayrı hücre üzerinden, test puanlarının yorumlanması ile kullanımını, ortaya konan kanıtlar ve ortaya çıkan sonuçlar açısından bir arada göstermiştir. Test uygulamasının dayanaklarını ve test uygulamasının işlev/çıktılarını, test uygulamasının iki ayrı yönü (facet) olarak tanımlayan Messick, dayanakları, kanıtlar ve sonuçlar temelinde ele almış; işlevler için ise test üzerinden yapılan yorumlar ve testin kullanımı olmak üzere iki başlık belirlemiştir (109-111). Test puanlarının yorumuna ait sütunda, ölçmeye ilişkin bilimsel değerlendirmelere yer verilmiş; testin kullanımına ait sütun ise testin kullanıldığı bağlama ilişkin değerlendirmelere ayrılmıştır. Matrisin orijinalinde hücrelerde numaralandırma bulunmamasına karşın, Messick tarafından matrise ilişkin olarak bir hücreden diğerine ilerleyişe ait yapılan açıklamanın, geçerliğin gösterilmesinde izlenen yolu da tarif etmesi nedeniyle, "progressive matris" bu numaralandırma ile birlikte kullanılmıştır (32). Matriste, yapı geçerliğinin her hücrede varlığına dikkat çeken Messick, yapı geçerliğini, test puanlarının anlamlandırılması, yorumlanması ve kullanılmasına ilişkin her türlü kanıtın entegrasyonu olarak ele almıştır (109-111).

Test puanları üzerinden yapılan yorumların farklı kanıtlarla bilimsel olarak değerlendirilmesi, yapı geçerliği olarak adlandırılmış (1 no'lu hücre); daha önceden kapsam geçerliği, ölçüt geçerliği gibi ayrı başlıklarda parçalı olarak kavramsallaştırılan geçerlikler, yapı geçerliği altında üniter bir anlayışla bu hücrede birleştirilmiştir. Test puanlarına ilişkin yorumlamaya bağlı ortaya çıkan sonuçlar ise testin altında yatan kuram, geçerliğe ilişkin tehditler ve test performasında neyin önemli/değerli olduğunun açıklanması ile ilişkilendirilmiştir (2 no'lu hücre).

Yapı geçerliğinin bir parçası olarak testin kullanıldığı bağlama göre uygunluğunun ve yararlılığının değerlendirilmesi, testin kullanımına dair kanıtlar üzerinden ele alınmıştır (3 no'lu hücre). Testin kullanımıyla ortaya çıkan sonuçlar ise testin kullanımının, bağlama özel, somut olarak ortaya çıkardığı sosyal sonuçlara ilişkin kanıtlara dayandırılmıştır (4 no'lu hücre).

Test puanlarının yorumlanmasına ilişkin "bilimsel" nitelikli sorular ile test puanları üzerinden yapılan yorumların kullanılmasına ilişkin "etik" niteliğe sahip soruların, yapı geçerliği temelinde nasıl entegre edileceğini göstermeyi amaçlayan matris, üniter geçerlik kavramsallaştırmasını da ortaya koymuştur. Anlaşılmasındaki zorluk, geçerliğin bilimsel ve etik boyutlarının ayrı ele alınmış olması ve aralarındaki ilişkinin açıkça gösterilememiş olması (112) gibi nedenlerle eleştirilen bu matris, 1990'lardan sonra Messick tarafından da fazla kullanılmamıştır (32).

1970'ler ve 1980'lerde Messick'in öncülük ettiği tartışmalarla şekillenen birleşik geçerlik

anlayışı egemen görüş haline gelerek geçerliğin gösterilmesi bir "bilimsel sorgulama" (scientific inquiry) yaklaşımıyla ele alınmış ve test puanlarının yorum ve kullanımlarının uygunluğu, yararlılığı, sosyal sonuçları ve kuramsal zemininin geçerlik kavramı kapsamına alınması genel olarak kabul görmüştür (16,32,62).

Messick'in üniter geçerlik anlayışının 1980'li yılların geçerlik kavramsallaştırması üzerindeki güçlü etkisine karşın, 1985 Standartları'nda üniter bir geçerlik tanımı yapılmamış ve Messick'in "bilimsel" ve "etik" boyutlarıyla ele aldığı geçerlik yaklaşımına Standartların bu güncellemesinde yer verilmemiştir (16,32,101). 1985 Standartlarında, önceki güncellemelerde olduğu gibi üç ayrı başlıkta kapsam, ölçüt ve yapıyla ilgili "kategoriler" belirtilmiş olmasına karşın, bu kategorilerle ilişkili "geçerlik kanıtlarından" söz edilmiş olması önemli bir dönüm noktası olarak kabul edilmiştir (16,32,101,103,108). 1985 Standartları, geçerliği, "test puanları üzerinden yapılan yorumların, kapsam, ölçüt ve yapıyla ilişkili ortaya konan kanıtlar üzerinden genel olarak değerlendirilmesi" olarak ele almış (104); geçerliğin gösterilmesinde üç geleneksel geçerlik kategorisiyle ilgili çeşitli kanıtların gösterilmesi gerekliliğini belirtmiştir (103).

Messick, 1980'lerin başında "progressive matrix" aracılığıyla açıkladığı birleşik geçerlik kavramsallaştırmasını 1990'lı yılların sonlarına kadar geliştirmiş, yapı geçerliği olarak ele aldığı geçerliğe ilişkin altı ayrı özellik tanımlamıştır. Geçerlikle ilgili temel konular zemininde organize ettiği bu özellikler ile geçerliğin farklı işlevsel yanlarını ortaya koyduğunu belirten Messick, test puanlarından yapılacak çıkarımların uygunluk, anlamlılık ve gerekliliğinin değerlendirilmesinde bu ayrımın yardımcı olacağını ifade etmiştir (109). Geçerliğin bu altı özelliğinin birbirinden ayrı ve birbiri yerine geçebilen farklı geçerlik tipleri olarak düşünülmemesi gerektiği ve geçerlik kanıtları arasındaki tamamlayıcı ilişkinin önemi Messick tarafından vurgulanmıştır (109). Messick'in geçerlik kuramına temel oluşturan ve geçerlik kanıtlarının ortaya konmasını gerektiren altı özellik şunlardır:

1- Kapsam açısından yapı geçerliği (The content aspects of construct validity): geçerliğin kapsamla ilişkili yanı için temel mesele, test ile ortaya çıkarılacak olan bilgi, beceri ya da diğer özelliklerin belirlenmesidir. Testin kapsamı, yapının ait olduğu alan ile uyumlu olmalı (relevance) ve onu yeterli biçimde temsil etmelidir (representativeness). Ölçülen yapıyla ilgili olmayan varyansa yol açacak zorluk ya da kolaylık gibi geçerlik tehditlerinin varlığına dikkat edilmelidir.

2- Yapı geçerliğinin öze ilişkin yanı (substantive aspect of construct validity): Ölçülen yapının gerektirdiği zihinsel süreç/işlemler ile test arasındaki uyumu vurgulayan bu özellik, zihinsel süreçlere ilişkin modelleme ya da yapıyı açıklayan temel kuramlar ile gösterilebilir. Konu alanının kapsam açısından temsiliyeti dışında, ölçülen özellikle ilişkili zihinsel süreçler için testin uygun örnekleme (sampling) sahip olması ve bu örneklemin uygunluğunun önceden yapılmış ampirik çalışmalarla gösterilmiş olması gereklidir.

3- Yapısal bakımdan geçerlik (structural aspect of validity): Ölçülmek istenen yapı ile test aracılığıyla gözlemlenen davranışlar arasındaki yapısal ilişkilerle Yapısal bakımdan geçerlik puanlama modeli kullanılması gerekir; test performansının puanlaması, ölçülen özelliğe bağlı olarak ortaya çıkan test davranışındaki değişiklikleri doğru biçimde yansıtmalıdır.

4- Yapı geçerliğinin genellenebilirliği (generalizability of construct validity): Ölçülmek istenen özelliğin ait olduğu alandaki kapsam ve süreçler test tarafından yeterli biçimde temsil ediliyorsa, test puanları üzerinden yapılan yorumlar da ölçülmek istenen yapı özelinde genellenebilir. Genellenebilirlik, yapıyı temsil eden maddeler/görevler ya da

farklı zaman ve bağlamlarda test puanlarından yapılan çıkarımlar arasındaki korelasyonun derecesine göre değişir.

5- Harici yanıyla yapı geçerliği (external aspect of construct validity): Bu özellik, test puanları üzerinden yapılan yorumlar ile ölçülen yapıyla ilgili olan diğer değişkenler arasındaki birleşen (convergent) ve ayrışan (discriminant) korelasyonlar temelinde araştırılmalıdır.

6- Sonuçlar bakımından yapı geçerliği (consequential aspect of construct validity): Geçerliğin sonuçlarla ilgili yanı, test puanlarının yorumlanması ve kullanılmasıyla ortaya çıkan istenen ve istenmeyen sonuçların değerlendirilmesi, kanıtların ve gerekçelerin gösterilmesiyle ilgilidir. Ortaya çıkan sonuçlar hem bireysel hem de grup üzerine etkileri açısından incelenmelidir.

1989 yılında yayınlanan "Educational Measurement" üçüncü baskısının geçerlik bölümü Messick tarafından hazırlanmış ve 1999 Standartları'nda da Messick'in görüşleri temel alınmıştır (101,103). 1999 Standartları'nda geçerlik, "testlerin tasarlandıkları kullanıma göre, değerlendirme puanları üzerinden yapılan yorumların teori ve kanıtlar tarafından desteklenme derecesi" olarak tanımlanmış (103); geçerliğin gösterilmesi (validation) için beş başlık belirlenmiştir. 1-Testin kapsamı temelindeki kanıtlar (evidence based on test content), 2- Yanıtlama süreçleri temelindeki kanıtlar (evidence based on response processes), 3- Testin içyapısı temelindeki kanıtlar (evidence based on internal structure), 4- Diğer değişkenlerle ilişkisi temelindeki kanıtlar (evidence based on relations to other variables), 5-Test uygulamasının sonuçları temelindeki kanıtlar (evidence based on consequences of testing) (32,101,103).

Messick'in geçerlik kuramına ilişkin temel özellikler Newton ve Shaw tarafından şu başlıklarda özetlenmiştir (32).

1-Geçerliğin gösterilmesinde söz konusu olan test değildir, test puanları üzerinden yapılan yorumların geçerliği söz konusudur. 2- Geçerlik kuramının merkezinde ölçüm yer almaktadır. 3- Geçerlik, yapı geçerliğidir ve yapı geçerliği puanlara verilen anlamla ilişkilidir. 4- Geçerlik, testin işlevleri ve geçerliğin gerekçelendirilmesi gibi farklı boyutlar içermektedir. 5- Geçerliğin gösterilmesi, deneysel, istatistiksel ve felsefi yollarla hipotezlerin ve bilimsel kuramların değerlendirildiği bilimsel bir çalışmadır. 6- Geçerliğin gösterilmesinde, geçerliğe ilişkin tehditler dikkate alınmalıdır. 7- Test puanlarının yorumlanmasına ilişkin "geçerlidir" ya da "geçersizdir" nitelemesi yapılamaz, geçerlik bir derece sorunudur. 8- Geçerliğin gösterilmesi, sonu olmayan, sürekli devam eden bir süreçtir.

Messick'in geçerlik kavramsallaştırmasının "bilimsel sorgulama" anlayışına dayanmış olmasından hareketle, Shepard bu anlayışa ilişkin iki konuya işaret etmiştir (16,62,112). İlk olarak, puanların anlamlandırılması (score meaning) üzerine yürütülen araştırmanın testin geçebileceğine ve testin geçerliğinin değerlendirilmesinde yanlış yönlendirmeye yol açabileceğine dikkat çekmiştir. Geçerliğin gösterilmesinin hiç bitmeyecek, sonu olmayan bir araştırma olduğu yaklaşımının, uygulayıcıları yıldırarak ne tür kanıt olursa olsun az sayıda kanıtla yeterliğin gösterilmesi riskinin ortaya çıkacağından söz etmiştir (16). Shepard ayrıca geçerliğin gösterilmesinde, uygulamanın bağlamına özel olarak geçerlikle ilgili öncelikli soruların ortaya konabilmesi ve sadece destekleyen değil, aynı zamanda geçerlik iddiasını çürüten kanıtların da gösterilmesi gereğinin üzerinde durmuştur (16,62,112). Shepard, Cronbach'ın argümana dayalı yaklaşımını ve Kane'in bu görüş üzerinde geliştirdiği modelin, geçerliğin gösterilmesi için uygun olduğunu savunmuştur (16,62).

5.5. Kane ve argümana dayalı geçerlik yaklaşımı

Messick tarafından öne sürülen birleşik geçerlik anlayışı tüm dünyada yaygın kabul görmekle

beraber, bilimsel ve etik konuları kapsayacak biçimde genişleyip büyüyen geçerlik kuramının, geçerliğin gösterilmesi için uygulamaya dönüştürülmesinde belirsizlikler doğmuştur (62). Diğer yandan, yapı geçerliğine ilişkin farklı felsefi yaklaşımlar çerçevesinde, tüm geçerliğin yapı geçerliği altında ele alınması da tartışma konusu olmuştur (32,62). Geçerliğin, uygulamada gerçekleştirilebilir parçalara ayrılarak ele alınması, kavramsal ve pratik olarak sadeleştirilmesi ihtiyacının ortaya çıktığı bu dönem, Newton ve Shaw tarafından geçerlik açısından "yapı bozumu" (deconstruction) dönemi olarak nitelendirilmiştir (32).

Yirmi birinci yüzyılla beraber geçerlik anlayışında köklü değişiklikleri içeren tartışmalar ve farklı görüşler ortaya çıkmıştır. 1970'lerden günümüze kadar geçerlik kuramlarının merkezinde rol alan "yapı" ve "yapı geçerliği"nin geçerlik kavramıyla ilişkisi, yapı geçerliğinin anlam ve önemi, tartışmaların temel konularından birisi olmuş; Borsboom, Markus, Lisstz, Samuelsen, Embretson ve Kane gibi geçerlik kuramcıları tarafından yapı geçerliği farklı açılardan sorgulanmaya başlanmıştır (58,101,102,104,113). Diğer bir tartışma, test sonuçlarının etik ve sosyal boyutlarının geçerlikle birlikte ele alınması konusunda geçerliğin kapsamına ilişkin anlayış farklılığına bağlı olarak ortaya çıkmıştır (11,61,103,112,113). Geçerlik kuramı çerçevesinde yürütülen tüm bu tartışmalar geçerliğin gösterilmesine ilişkin uygulama açısından da belirleyici olmuştur. Cronbach tarafından 1980'lerde tartışmaya açılan argümana dayalı yaklaşım, yöneltilen farklı eleştiriler olmakla beraber, bu dönemde öne çıkan geçerlik yaklaşımı olmuştur (32).

Cronbach'ın 1980'li yıllarda geçerliğin gösterilmesine ilişkin "geçerlik araştırması" (validation research) yerine "geçerlik argümanı" (validiy arguement) yaklaşımına dikkat çekmesi ile gündeme gelen "argümana dayalı geçerlik" kavramı, Kane tarafından geliştirilerek yeni bir geçerlik yaklaşımına dönüşmüştür (17,32,104,115). Kane, üniter geçerlik modelini kavramsal olarak zengin ve çekici bulduğunu belirtmekle beraber, geçerliğin gösterilmesinde nereden başlanacağı, nasıl ve nereye kadar ilerleneceği, ne zaman durulacağına ilişkin bir yol göstermemesi nedeniyle eleştirdiği modelin uygulanmasının kolay olmadığını ifade etmiştir (115). Soyutluğu nedeniyle uygulamada sorunlara yol açtığını belirttiği Messick'in yaklaşımının bu sınırlılıklarına karşı, geçerliğin gösterilmesi için, argümana dayalı yaklaşımı öne sürmüştür (17,104,115). Kane'in 1990'lı yıllarda öne sürdüğü bu yaklaşım, 2006'da Educational Measurement dördüncü baskısında, Kane tarafından hazırlanan "Test Validation" bölümünde, yer almıştır (3). 1999 Standartları'nda beş başlık altında ele alınan geçerlik kanıtlarına ilişkin kategoriler, 2014 Standartları'nda devam etmekle beraber, son güncelleme, argümana dayalı geçerlik yaklaşımını benimsemiştir (116).

Kane, 1999 Standartlarında yer alan geçerlik anlayışı ile kendi geçerlik yaklaşımının ortak temel özelliklere sahip olduğunu belirtmiştir (17). Geçerliğin, test puanlarının yorum ve kullanımıyla ilişkili bir konu olarak ele alınması gerekliliği Kane tarafından da belirtilmiştir. Borsboom ve Mellengbergh, Messick ve Kane'in görüşlerinden farklı olarak, geçerliğin test sonuçlarının yorumlanması ve kullanımına ait bir özellik olmayıp, ölçüm aracına ait bir özellik olduğunu; ölçülmesi hedeflenen özelliğe ait değişmelere gösterdiği duyarlığa göre ölçme aracının geçerliğinin değerlendirmesini savunmaya devam etmişlerdir (32,61,91,113).

Geçerliğin, test puanlarına ilişkin yorumlara temel oluşturan çıkarımların uygunluk ve yeterliliğinin kuramsal dayanak ve ampirik kanıtlar tarafından desteklenme derecesine ilişkin entegre bir değerlendirme olarak görülmesi, Messick ve Kane'in yaklaşımlarındaki diğer bir ortak özelliktir (17). Bunun dışında, Messick'in görüşlerine uygun biçimde, Kane tarafından

öne sürülen argümana dayalı yaklaşım da, test uygulamasına bağlı olarak ortaya çıkan sonuçların tartışılmasını geçerliğin gösterilmesi için önemli bulmuştur (17,104,115).

Argümana dayalı yaklaşım, testin cevaplanmasından, test puanlarının yorum ve kullanımına kadar olan aşamaların her birinde belli çıkarım ve varsayımlar üzerinden hareket edildiğini temel almıştır. Kane, geçerliğin gösterilmesini kolaylaştırmak amacıyla teste ait çıkarım ve varsayımların kullanıldığı dört temel aşama tanımlamış; puanlamadan karar vermeye kadar olan tüm aşamaları, çıkarım ve varsayımlar zinciri olarak nitelemiştir (17,104,115).

1- Puanlama çıkarımları (scoring inferences) - gözlemden gözlemlenmiş puana: Puanlama çıkarımları, testteki her bir maddeye verilen yanıtlar için atanacak puan değeri ve teste ait toplam puanın belirlenmesi ile ilgili çıkarımlardır. Testin niteliği, testin uygulanması ve puanlamayla ilgili belli özellik ve süreçleri (ölçülecek özellik ile testte yer alan soruların ilişkisi, testin süresi, uygulanma biçimi ve ortamı, maddelere ilişkin atanacak puan değerleri, toplam puanın hesaplanması, vb) kapsayan bu aşamada kullanılan "puanlama çıkarımları", puanlama işlemlerinin uygun olduğu, planlandığı gibi uygulandığı ve yanlılık (bias) içermediği, ölçülecek özellikle ilgisi olmayan varyansı (construct irrelevant variances) içermediği gibi temel varsayımlara dayanır. Puanlama çıkarımlarının dayandığı varsayımların kabul edilebilirliği, bu aşamaya ait işlem ve süreçlerin tanımlanmasını; test puanlarında görülen değişikliğin, ölçülmesi hedeflenen özelliğe bağlı olarak ortaya çıktığına dair kanıtların ortaya konmasını gerektirir.

2- Genelleme çıkarımı (generalization inference) – gözlemlenmiş puandan evrensel puana: Genelleme çıkarımı, testten elde edilen toplam puanın, benzer şartlarda benzer testlerde beklenen performans için bir gösterge olabileceğine, genellenebileceğine dair çıkarımdır. Testte yer alan soruların nitelik ve sayısının, testin uygulanma biçimi ve ölçme hatasını azaltacak özellikte olduğu gibi varsayımlara dayanır. Genelleme çıkarımlarının kabul edilebilirliği için güvenirlik ve genellenebilirlik çalışmaları, ölçmenin standart hatası, standart sapma gibi değerlerin dikkate alınması gereklidir.

3- Kestirim çıkarımları (extrapolation inferences) – evrensel puandan hedef alana: Kestirim çıkarımları, testin genellenebilir puanlarından, hedef alandaki gerçek yaşamda gösterilecek performansa ilişkin yapılan çıkarımlardır. Bu aşamada, puanların farklılığı değil, puanlara atfedilen anlam yorumlara ilişkin farklılıklar ele alınır ve gelecekteki performansa ilişkin çıkarım yapılır. Kestirim çıkarımları, gerçek performansın göstergesi niteliğindedir ve test kapsamının, testte yer alan maddelerin dağılımının gerçek yaşamdaki performansı kestirmeye uygun olduğu varsayımına dayanır. Testte yer alan maddelerin ve testin hazırlanma süreçlerinin incelenmesi, kapsam analizi, başka ölçütle korelasyonun araştırılması, yüksek sesle düşünme protokolleri (think aloud protocols) gibi farklı kanıtlar aracılığıyla yorum ve kullanım özelliklerine göre bu varsayımların kabul edilebilirliği değerlendirilir.

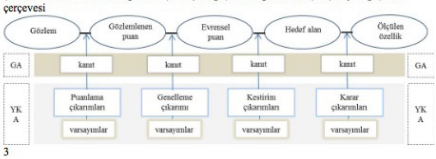
4- Karar çıkarımı (decision inference) – hedef alandan ölçülen özelliğe: Test sonuçları üzerinden yapılan yorumlar, bireyler ya da gruplar hakkında karar vermede kullanılır; karar çıkarımı da, söz konusu kararlara esas oluşturan kurallara ilişkindir. Bu kuralların, test puanlarının kullanımını takiben ortaya çıkması beklenen sonuçlara yol açacak biçimde iş gördüğü çıkarımı, karar çıkarımı olarak nitelendirilir. Karar çıkarımı, karar vermede uygulanan kuralların beklenen olumlu sonuçlara yol açacağı ve istenmeyen sonuçların ortaya çıkmayacağı varsayımına dayanır. Bu varsayımların kabul edilebilirliği ise, testin uygulandığı bağlam ve grup için, test uygulamasının beklenen sonuçlarına ne

derecede ulaşıldığının, testin uygulandığı grup ve diğer gruplar üzerinde farklı etkilerinin neler olduğunun, özellikle eğitim alanında kullanılan testlerin eğitim sistemi üzerindeki olumlu ve olumsuz etkilerinin araştırılmasını gerektirir. Ortaya çıkan olumlu ve olumsuz sonuçlara ilişkin kanıtlar doğrultusunda karar çıkarımı için öne sürülen varsayımın kabul edilebilirliği değerlendirilir.

"Yorum/kullanım argümanı" ve "geçerlik argümanı" Kane tarafından, argüman temelli geçerlik yaklaşımının iki temel bileşeni olarak tanımlanmıştır (17,104,115,117). Test puanlarının yorumlanması ve bu yoruma dayalı kullanımıyla ilişkili olarak her test uygulamasının kendi bağlamında açık ve net biçimde ifade edilmesi gerekli olan çıkarım ve varsayımların tamamı, testin "yorum/kullanım argümanı" (YKA) (interpretation/use arguments-IUAs) olarak adlandırılmıştır (17,104,115). YKA, geçerlik kanıtlarının toplanması ve sunumu için gereken çerçeveyi oluşturur; YKA'yı oluşturan farklı çıkarım ve varsayımları destekleyen ve çürüten farklı özellikteki kanıtlar üzerinden YKA'nın uygunluk ve yeterliğinin değerlendirilmesi Kane tarafından "geçerlik argümanı" (validity argument) olarak nitelendirilmiştir (17,104,115). Kane, YKA ve geçerlik argümanlarının tanımlanma ve değerlendirilmesini kapsayan bu süreci, geçerliğin gösterilmesi için bir metodoloji olarak öne sürmüştür (17,104). Argümana dayalı geçerlik yaklaşımın genel çerçevesi aşağıda verilmiştir.

(Şekil -2)

Şekil 2: Yorum kullanım argümanı (YKA) ve geçerlik argümanına (GA) dayalı geçerlik çerçevesi



3

Kane, test uygulamasının bağlam ve kullanım özelliklerine bağlı olarak, bazı test uygulamaları için belli zincirlerin çıkarım ve varsayımlarının öne çıkıp, bunlarla ilgili kanıtların gösterilmesinin ağırlık ve önem kazanacağını; diğer bazı test uygulamaları için başka zincirlerin çıkarım ve varsayımları üzerinden farklı kanıtların ortaya konması gerekeceğini belirtmiştir (104,115). Argümana dayalı yaklaşımda, YKA'yı destekleyen ve çürüten kanıtların bir arada ele alınmasıyla en zayıf halkanın belirlenmesine dikkat çekilmiştir (17,104,115).

Önceki geçerlik anlayışından farklı olarak, Kane, geçerliğin gösterilmesinin, bilimsel nitelikle sınırlı, sonu olmayan, hiç bitmeyecek bir araştırma olarak ele alınmasına itiraz etmiştir (104). Geçerliğin değerlendirilmesi için YKA ve geçerlik argümanı çerçevesinde Kane tarafından önerilen metodolojinin, bağlama özel, uygulamaya dönük, esnek ve pragmatik bir yaklaşım olduğu belirtilmiştir (16,32,98,104). Moss, Kane'in geçerlik anlayışı ile Messick'in geçerlik kuramının birbiriyle uyumlu olduğunu belirtmiş, ancak Messick'in "bilimsel sorgulama" olarak nitelendirdiği geçerliğin, Kane tarafından "practical argument" anlayışıyla ele alındığına işaret etmiştir (16).

Kane tarafından öne sürülen geçerlik anlayışının önceki geçerlik kuramlarından diğer bir farkı, geçerliğin gösterilmesi için tek bir "yapı" kavramı kullanmamış olmasıdır (32). Kane, ölçmeye konu olan tek tip bir yapı yerine, "kuramsal yapılar" (theoretical constructs-indicators) ve "gözlemlenebilen özellikler" (observable attributes) olarak ayrım yapmıştır (17,104). Kane, kuramsal zeminde yorumlanan yapıların, ilişkili oldukları kuramdaki rolleri üzerinden tanımlandığını, "gözlemlenebilen özelliklerin" ise testin ait olduğu alanda yapılabilecek gözlemler evrenine göre tanımlandığına dikkat çekmiştir (104). Test puanlarına ilişkin yorumların niteliğinin ve geçerlik için ortaya konması gereken kanıtların bu ayrıma bağlı olarak değişeceğini belirten Kane, "yapı" temelinde kurama dayalı yapılacak

yorumların nedensellik ilişkisini gösterecek biçimde kuramsal ve ampirik nitelikte daha fazla kanıt gerektirdiğini; oysa nedensellik ilişkisinin aranmadığı "gözlemlenebilen özellikler" için açıklayıcı bir modelin yeterli olduğunu belirtmiştir (104). Yapı geçerliğini temel alan üniter geçerlik yaklaşımında ve parçalı modelin yapı geçerliğinde, ölçülen yapıya ilişkin bu tür bir ayrım yapılmamış, yorumlar ve kanıtlar standart bir çerçevede ele alınmıştır (32). Özellikle eğitim alanında kullanılan başarı testleri, beceri sınavları, performans değerlendirmelerinde geçerliğin gösterilmesi ve test puanlarının yorumlanmasıyla, kuramsal ilişkilendirme gerektiren daha karmaşık psikolojik yapılarla ilgili geçerlik çalışmaları arasındaki fark Kane tarafından örnekleriyle açıklanmıştır (104,115).

Kane'in yaklaşımı, her test uygulaması için geçerlik kanıtlarına ilişkin standart bir model tanımlamamıştır (16,17,32,117). Kane, YKA'nın, testin uygulandığı bağlama göre bütünüyle değişebileceğini, ortaya konacak kanıtlar için de bu değişikliğe uygun, esnek bir yol izlenmesi gerektiğini belirtmiş; önceden düzenlenmiş, sabit bir çerçeveye uyma zorunluluğu yerine geçerliğin gösterilmesi için test uygulamasına özel yorum ve kullanımın dikkate alınması gerektiğini savunmuştur (104,115). Test uygulamasının gerektirdiği bağlama özel, esnek bir yol izlenebilmesine fırsat veren bu yaklaşım, validasyon sürecini evrensel ve standart biçimde tanımlayan diğer geçerlik anlayışlarından farklı olarak, geçerliğin gösterilmesinde her testin kendi bağlamına göre şekillenecek varsayım, çıkarımlara dayalı yorum kullanım argümanını temel almıştır (17,32,98,104).

Kane'in argümana dayalı geçerlik yaklaşımı, 2000'li yıllar boyunca kabul gören yaklaşım olmasına karşın eğitim alanında geçerlik konusunda farklı görüşler öne sürülmüştür. Messick ve Kane'in geçerlik yaklaşımlarından hareketle, geliştirici değerlendirme (formative assessment) ve aralıklı değerlendirme (interim assessment) bağlamına göre değiştirilmiş geçerlik çerçeveleri bu bağlamda ele alınabilir (118,119).

1980'lerden sonra eğitim amaçlı değerlendirme ile psikoloji alanındaki ölçmenin birbiri ile çok benzer olmadığı görüşü (66) doğrultusunda hem değerlendirme uygulamaları hem de geçerlik yaklaşımları değişmeye başlamıştır. Eğitim alanında, 1980'lere kadar, sınav puanlarına odaklı, "ölçme temelli" geçerlik yaklaşımının egemen olduğu farklı yazarlar tarafından vurgulanmıştır (13,67,118). 1990'ların başından itibaren, değerlendirme yaklaşımına ilişkin değişimin ortaya çıkmasıyla beraber, geçerlik kavramının ele alınışında da önemli tartışmalar yürütülmüştür (15,17,61,62,93,99). Ölçmenin konusu olan "yapı", ölçmeye dayanak oluşturan varsayımlar, ölçüm sonuçlarının yorumlanması ve kullanılması gibi konuların geçerlik kapsamında tartışılması geçerlik tanımını ve geçerlik ile ilgili kavramsallaştırmayı ciddi biçimde etkilemiş ve değiştirmiştir (16,93,99,101).

Brookhart, eğitim sürecinde öğrencilerin öğrendiklerinin kapsam ve niteliğine ilişkin bilginin edinilmesi ve eğitsel kararların alınması amacıyla eğiticiler tarafından yapılan değerlendirmelerin, geniş ölçekli, standardize testlerden amaç ve işlev olarak farkını vurgulayarak, psikometrik anlayışa dayalı bu tür testler bağlamında ele alınan geçerlik, güvenirlik kavramlarının eğiticilerin kendi değerlendirmelerine her zaman hizmet etmeyeceğini belirtmiştir (44). Geniş ölçekli testlerin kullanıcılarının yönetici ve politika yapıcılar olduğuna dikkati çeken Brookhart, değerlendirmenin kullanıcılarının öğrenci, eğitici ve okulun kendisi olduğunu ifade etmiştir (44).

Geçerlik kuramı çerçevesinde öne sürülen anlayışların ölçüm temelli değerlendirme uygulamalarını dikkate aldığını belirten Moss, eğiticilerin günlük değerlendirme uygulamalarını ve geliştirici değerlendirmeyi

(formative assessment) kapsayacak biçimde, değerlendirmeyi sadece teknik değil, sosyal bir süreç olarak ele alan bir yaklaşımla geçerlik kuramının değişmesi ve gelişmesi gerektiğini ifade etmiştir (16,93). Moss ve arkadaşları, ölçme ile sınırlı olmayan bir değerlendirme anlayışı çerçevesinde; değerlendirmenin eğitici, öğrenciler ve öğrenme ortamı arasındaki etkileşimli rolünü ortaya çıkaracak olan olgu çalışmaları ile bağlama özel geçerliğin araştırılmasının önemini vurgulamışlardır. Aynı yazarlar, "scientific inquiry", "practical argument" yaklaşımlarından farklı olarak, sosyal bilimlerin yorumlayıcı geleneğinden yararlanan "situated inquiry" yaklaşımına dayalı geçerlik kavramsallaştırmasını öne sürmüşlerdir (16).

Sonuç:

Değerlendirme ve geçerlik konularına ilişkin tarihsel süreç dikkate alındığında, iki temel saptama yapılabilir. Birincisi, ölçme kuramlarına dayalı değerlendirme ve geçerlik yaklaşımı ve bu yaklaşımların altında yatan temel varsayımlar, 20. yüzyıl ilk yarısının sosyal, politik, ekonomik ortam ve koşullarında, dönemin egemen paradigması olan pozitivizm temelinde ortaya çıkmıştır. İkincisi, 20. yüzyılın ikinci yarısından sonra bu yaklaşım ve varsayımlar sorgulanmış, farklı değerlendirme yaklaşımları üzerinde yürütülen tartışmalarla beraber, değerlendirmenin "teknik bir iş" olduğu görüşü eleştirilerek ölçmeye dayalı değerlendirme anlayışı değişmeye başlamıştır. Ancak tüm eğitim kademelerinde olduğu gibi yükseköğretimde de, değerlendirme uygulamaları, ağırlıklı olarak ölçme kuramlarına dayalı "bilimsel söylem"" çerçevesinde planlanıp, uygulanmaya ve tartışılmaya devam edilmektedir.

Delandshere'e göre, öğrenmeye ve değerlendirmeye ilişkin anlayış değişikliklerine karşın, ölçme kuramlarına dayalı söylem ve uygulamaların pratikte egemenliğini sürdürdürmesi, eğitim dünyasının bu söylem ve uygulamaların doğruluğu üzerinde görüş birliğine sahip olmasından çok, yeni nesil eğitimciler, değerlendirme uzmanları ve sınavların farklı kullanıcılarının, nereden kaynaklandığını ve nasıl geliştiğini gerçekten bilmeden ve sorgulamadan bu geleneğe ait yöntem, kavram ve standartları kullanarak, aynı geleneğin içinde yer almalarından kaynaklanmaktadır (23). Bu durumda, profesyonel uygulamalara ilişkin "reflektif uygulama"nın önemi ortaya çıkmaktadır.

"Teknik rasyonalite" (technical rationality) ve "reflektif uygulama" (reflective practice) olarak iki tür profesyonel yaklaşımdan söz eden Schön, bu iki yaklaşım arasındaki önemli farka dikkat çekmiştir (30). Schön'e göre, teknik rasyonalite, karşılaşılan durum ya da problemin, uzmanlık alanının "bilimsel bilgi" ve kuralları üzerinden ele alınarak, durum ya da probleme uygun teknik, yöntem, araçların seçimiyle bilimsel bilginin uygulamaya geçirilmesidir (30). Reflektif uygulama ise sahip olunan bilimsel bilginin doğrudan uygulamaya transferinden çok, karşılaşılan her durum ve problemde kendi deneyimlerimizi, konuya ilişkin anlayışımızı ve duruma ilişkin bağlamı sorgulamamızı da gerektirir (30). Profesyonel uygulamada karşılaşılan sorunların, büyük ölçüde, teknik rasyonel yaklaşımla çözülebilecek net tanımlanmış, sabit ve düzenli bir zeminde ortaya çıkan teknik sorunlar olmadığını vurgulayan Schön, profesyonel uygulama alanının, bataklık alanlardaki düzensizlik ve karışıklığa benzer biçimde, belirsiz, değişken, çatışmalı bir zemine sahip olduğunu ve reflektif uygulama gerektirdiğini belirtmiştir (30).

Schön'ün kavramsallaştırdığı reflektif uygulama yaklaşımının yanı sıra, Mezirow'un öncülüğünde 1990'lardan itibaren eğitim alanındaki önemiyle tartışılmaya başlanan eleştirel refleksiyon da, farklı disiplin ve mesleklerin profesyonel uygulamaları için temel bir nitelik olarak ele alınmaktadır (120). Mezirow eleştirel refleksiyonu, bulunduğumuz

bağlamda edindiğimiz bilgi, anlayış ve inançlarımızın uygunluğunu değerlendirmek, doğru kabul ettiklerimizin geçerliğini analiz etmek ve sorgulamak, bakış açımızı ve neden öyle düşünüp davrandığımızı gözden geçirmek ve bakış açımızı geliştirip değiştirebilmek olarak tanımlamıştır (121). Fook, White ve Gardner'a göre, eleştirel refleksiyon, bireylerin eylemlerini yönlendiren temel varsayımların tarihsel ve kültürel kökenleriyle beraber tanımlanabilmesi, farkına varılması, bu varsayımların sorgulanması ve alternatif uygulamaların neler olabileceğine ilişkin yaklaşımın geliştirilmesini gerektirir (122).

Hekimlik açısından reflektif uygulama ve eleştirel refleksiyonun önemi, tıp eğitimi alanında farklı yazarlar tarafından vurgulanmış, mezuniyet öncesi ve sonrası tıp eğitimi ve sürekli mesleksel gelişim bağlamında reflektif uygulamanın geliştirilmesine yönelik strateji, yöntem ve teknikler tartışılmıştır (123-128). Tıp Eğitimi disiplini açısından reflektif uygulama ve eleştirel refleksiyon, sadece hekimlik bağlamında tıp eğitimi sürecinde kazandırılması gereken bir nitelik olması açısından değil, her disiplin için olduğu gibi alanın kendi profesyonel uygulamaları ve bilimsel çalışmalar açısından da önemlidir. Sosyal, politik, kültürel, ideolojik, kurumsal dinamikler çerçevesinde açık bir sistem olan eğitim alanındaki profesyonel uygulamaların "karmaşık ve düzensiz zemini", alanın ihtiyaçları ve sorunların çözümünde teknik rasyonel bir yaklaşımı genellikle geçersiz kılmaktadır. Diğer yandan, profesyonel gelişim, değişim ve dönüşüm, strateji, teknik ve yöntemlere ilişkin daha çok bilgi ve daha çok tecrübe sahibi olmanın ötesinde, bildiklerimizin, düşüncelerimizin, kararlarımızın geçerliğini gözden geçirme, düşünce ve eylemlerimizi belirleyen bilimsel söylemi sorgulama çabası ile birlikte mümkün olabilir.

Tıp eğitimi ile ilgili tüm konularda olduğu gibi değerlendirme konusunda da, egemen bilimsel söylemi ve farklı yaklaşımları, tıp eğitimi bağlamıyla sınırlı kalmadan ama tıp eğitimi bağlamını dikkate alarak sorgulayıp, kendi söylemimizi, uygulamalarımızın neden ve sonuçlarını gözden geçirip kendimize yeni sorular sorabiliriz. Örneğin, öğrenme ve öğrencilere ilişkin değerlendirmeyle ilgili olarak günlük söylemimizde, yayın ve resmi dokümanlarda neden "ölçme değerlendirme" ifadesini kullandığımız sorusu, bu sorulardan birisi olabilir.

Kavram ve anlayışların değiştiği günümüzde, değerlendirme teknik ve yöntemleriyle ilgili yöneltilen sorulara vereceğimiz cevaplardan çok, öğrenme ve değerlendirme bağlamında kendi epistemolojik inançlarımıza, değerlendirmeye ilişkin kullanılan "bilimsel söyleme" ve bildiklerimizin neye göre ve neden doğru olduğuna dair soracağımız sorulara daha fazla ihtiyacımız olabilir. Belki de, soracağımız daha çok soru, vereceğimiz daha az cevap vardır.

Kaynaklar

1. Demirel Ö. Eğitim Sözlüğü. Üçüncü Baskı. Ankara: Pegem A Yayıncılık; 2005.

2. Atılgan H, Kan A, Doğan N. Eğitimde Ölçme ve Değerlendirme. İkinci Baskı. Ankara: Anı Yayıncılık; 2007.

3. Tekin H. Eğitimde Ölçme ve Değerlendirme. On üçüncü Baskı. Ankara: Yargı Yayınları; 1991.

4. Turgut MF. Eğitimde Ölçme ve Değerlendirme Metotları. Onuncu Baskı. Ankara: Yargıcı Matbaası; 1983.

5. Baykul Y. Eğitimde ve Psikolojide Ölçme: Klasik Test Teorisi ve Uygulaması Ankara: ÖSYM Yayınları; 2000.

6. Michell J. Measurement in Psychology: A Critical History of a Methodological Concept

(Ideas in Context) New York: Cambridge University Press; 1999.

7. Linn RL, Miller DM. Measurement and Assessment in Teaching. Ninth Edition ed. New Jersey: Pearson Education, Inc.; 2005.

8. Knight P. Grading, classifying and future learning. In Boud , Falchikov , editors. Rethinking Assessment in Higher Education. New York: Routledge; 2007. p. 72-86.

9. Dixon-Román, Ezekiel J; Gergen, Kenneth J. Epistemology and Measurement: Paradigms and Practices I. A Critical Perspective on the Sciences of Measurement. Princeton NJ: The Gordon Commission on the Future of Assessment in Education; 2012.

10. Gipps C. Beyond Testing. Towards a Theory of Educational Assessment London: The Falmer Press; 1994.

11. Moss PA, Pullin D, Gee , Haertel EH. The idea of testing: psychometric and sociocultural perspectives. Measurement: Interdisciplinary Research and Perspectives. 2005; 3(2): p. 63-83.

12. Black P, Dylan W. Assessment and classroom learning. Assessment in Education: Principles, Policy & Practice. 1998; 5(1): p. 7-74.

13. Delandshere G, Petrosky AR. Assessment of complex performances: limitations of key measurement assumptions. Educational Researcher. 1998; 27(2): p. 14-24.

14. Pellegrino JW, Chudowsky N, Glaser R. Knowing What Students Know: The Science and Design of Educational Assessment. Washington DC: Committee on the Foundations of Assessment, Center for Education, National Research Council, Rethinking the Foundations of Assessment; 2001.

15. Newton PE. Clarifying the consensus definition of validity. Measurement: Interdisiciplinary Research and Perspectives. 2012; 10(1-2): p. 1-29.

16. Moss PA, Girard BJ, Haniford LC. Validity in educational assessment. Review of Research in Education. 2006; 30(1): p. 109-162.

17. Kane MT. Current concerns in validity theory. Journal of Educational Measurement. 2001; 38(4): p. 319-342.

18. Foucault M. Bilginin Arkeolojisi İstanbul: Ayrıntı Yayınları; 2011.

19. Kuper A, Whitehead C, Hodges BD. Looking back to move forward: Using history, discourse and text in medical education research: AMEE Guide No. 73. Medical Education. 2013; 35: p. e849-e860.

20. Rehm J. The Use of Foucault in the Creation of Educational History: A Review of Literature. In Plakhotnik MS, Nielsen SM, Pane DM, editors. Proceedings of the 11th Annual College of Education & GSN Research Conference; 2012; Miami. p. 150-157.

21. Ramirez O, Boli J. The political construction of mass schooling: European origins and worldwide institutionalization. Sociology of Education. 1987; 60(1): p. 2-17.

22. Willbrinck B. Assessment in historical perspective. Studies in Educational Evaluation. 1997; 23(1): p. 31-48.

23. Delandshere G. Implicit theories, unexamined assumptions and the status quo of educational assessment. Assessment in Education: Principles, Policy & Practice. 2001; 8(2): p. 113-133.

24. Madaus GF, O'Dwyer LM. Short history of

performance assessment: lessons learned. Phi Delta Kappan. 1999; 80(9).

25. Meroe AS. Democracy, Meritocracy and the Uses of Education. Princeton, NJ: The Gordon Commission on the Future of Assessment in Education; 2012.

26. Stray C. The shift from oral to written examination: Cambridge and Oxford 1700–1900. Assessment in Education: Principles, Policy & Pract. 2001; 8(1): p. 33-50.

27. Cheetham G, Chivers G. Professions, Competence and Informal Learning Cheltenham: Edward ElgarPublishing; 2005.

28. Gipps C. Socio-cultural aspects of assessment. Review of Research in Education. 1999; 24: p. 355-392.

29. Sutherland. Examinations and the construction of professional identity: A case study of England 1800–1950. Assessment in Education: Principles, Policy & Practice. 2001; 8(1): p. 51-64.
30. Schön DA. The Reflective Practitioner: How Professionals Think in Action New York: Basic Books; 1983.

31. Hoskin KW, Macve RH. Accounting and the examination: A genealogy of disciplinary power. Accounting Organizations and Society. 1986; 11(2): p. 105-136.

32. Newton PE, Shaw SD. Validity in Educational&Psychological Assessment. 1st ed. London: Sage Publications Ltd; 2014.

33. Kaestle C. Testing policy in the United States: A historical perspective. Princeton, NJ: Gordon Commission on the Future of Assessment in Education; 2012.

34. Glaser R, Silver E. Assessment,Testing, and Instruction: Retrospect and Prospect. CSE Technical Report. Washington DC: National Center for Research on Evaluation, Standards and Student Testing; 1994.

35. Madaus GF, Stufflebeam DL. Program Evaluation: A Historical Overwiew. In Stufflebeam DL, Madaus GF, Kellaghan T, editors. Evaluation Models: Viewpoints on Educational and Human Services Evaluation. Boston: Kluwer; 2000. p. 3-18.

36. Shepard LA. The role of assessment in a learning culture. Educational Researcher. 2000; 29(7): p. 4-14.

37. Mershon S, Schlossman S. Education, science, and the politics of knowledge: The American Educational Research Association, 1915–1940. American Journal of Education. 2008;(114): p. 307–340.

38. Koppes LL, Pickren W. Industrial and Organizational Psychology: An Evolving Science and Practice. In Koppes LL, editor. Historical perspectives in industrial and organizational psychology. Mahwah, NJ: Lawrence Erlbaum Associates; 2007. p. 3-36.

39. McArthur L. Educational Testing and Measurement: A Brief History. CSE Report No:216. Los Angeles: National Institute of Education; 1983.

40. McGaghie WC. Assessing readiness for medical education: evolution of the Medical College Admission Test. JAMA. 2002;(288): p. 1085–90.

41. Melnick DE, Dillon GF, Swanson DB. Medical Licensing Examinations in the United States. Journal of Dental Education. 2002;(66): p. 595–9.

42. Filer A. Technologies of Testing: Editor's

Introduction. In Filer A, editor. Assessment : Social Practice and Social Product. London: Routledge Falmer; 2000. p. 43.

43. Dragositz A. The National Council on Measurement in Education: Its History, Purposes and Activities. In Yearbook of the National Council on Measurement in Education No. 20.: National Council on Measurement in Education; 1963. p. 170-172.

44. Brookhart S. Developing measurement theory for classroom assessment purposes and uses. Educational Measurement: Issues and Practice. 2003; 22(4): p. 5-12.

45. Ertürk R. Modern ve Postmodern düşüncelerde bilim. Felsefe Dünyası. 2004; 2(40): p. 65-76.

46. Pauli HG, White KL, McWhinney IR. Medical education, research, and scientific thinking in the 21st century (Part One of Three. Education for Health. 2000; 13(1): p. 15-25.

47. Michell J. Quantitative science and the definition of measurement in psychology. British Journal of Pyhology. 1997; 88: p. 355-383.

48. Mills JA. Operationism, Scientism and the Rhetoric of Power. In Tolman CW, editor. Problems, Positivism in Psychology: Historical and Contemporary. New York: Springer Verlag; 1991. p. 67-82.

49. Borsboom D. The attack of the psychometricians. Psychometrica. 2006; 71(3): p. 425-440.

50. Bechtel W, Adele A, Graham G. The Life of Cognitve Science. In Bechtel W, Graham G, editors. A Companion to Cognitive Science. Massachuetts, USA: Blackwell Publishers; 1998.

51. Pellegrino JW, Baxter GP, Glaser R. Addressing the "Two Disciplines" problem: linking theories of cognition and learning with assessment and instructional practice. Review of Research in Education. 1999; 24: p. 307-353.

52. Carver RP. Two dimensions of tests: psychometric and edumetric. American Psychologist. 1974; 29(7): p. 512-518.

53. Vinchur AJ. A History of Psychology Applied to Employee Selection. In Koppes LL, editor. Historical Perspectives in Industrial and Organizational Psychology. New York: Pschology Press; 2014. p. 193-218.

54. Cronbach LJ. My current thoughts on coefficient alpha and successor procedures. Educational and Psychological Measurement. 2004; 64(3): p. 398-418.

55. Gregory RJ. The History of Psychological Testing. In Psychological Testing: History, Principles,and Applications. Fourth Edition ed. : Allyn & Bacon; 2004. p. 1-28.

56. Jones LV, Thissen D. A History and Overview of Psychometrics. Handbook of Statistics. 2007; 26: p. 1-27.

57. Goldstein H. Francis Galton, measurement, psychometrics and social progress. Assessment in Education: Principles, Policy & Practice. 2012; 19(2): p. 147-158.

58. Embretson SL. The Second Century of Ability Testing: Some Predictions and Speculations. Princeton, NJ: Educational Testing Service; 2003.

59. Traub RE. Classical test theory in historical perspective. Educational Measurement Issues and Practice. 1997; 16: p. 8-14.

60. Michell J. Is psychometrics pathological

science? Measurement: Interdisciplinary Research and Perspectives. 2008; 6(1-2): p. 7-24.

61. Borsboom D. Measuring The Mind Cambridge: Cambridge University Press; 2005.

62. Shepard LA. Evaluating test validity. Review of Reseacrh in Education. 1993; 1(9): p. 405-450.

63. Hattie J, Jaeger RM, Bond L. Persistent methodological questions. Review of Research in Education. 1999; 24: p. 393-446.

64. Biggs J, Tang J. Teaching for Quality Learning in University. 3rd ed. Berkshire: Open University Press& McGraw-Hill Companies; 2007.

65. Falchikov N. The place of peers in learning and assessment. In Boud D, Falchikov N, editors. Rethinking Assessment in Higher Education: Learning for the long term. New York: Routledge; 2007. p. 128-143.

66. Baird JA, Black p. Test theories, educational priorities and reliability of public examinations in England. Research Papers in Education. 2013; 28(1): p. 5-21.

67. Baird JA, Hopfenbeck TN, Newton P, Stobart G, Steen-Utheim AT. Assessment and Learning: State of the field review. Oxford University Centre for Educational Assessment Report OUCEA/14/2, Valuable Learning; 2014.

68. McCourt W. Paradigms and their development: The psychometric paradigm of personnel selection as a case study of paradigm diversity and consensus. Organization Studies. 1999; 20(6): p. 1011-1033.

69. Driskell JE, Olmstead. Psychology and the military. American Psychologist. 1989; 44(1): p. 43-54.

70. Garavan TN, McGuire D. Competencies and workplace learning: some reflections on the rhetoric and the reality. Journal of Workplace Learning. 2001; 13(4): p. 144-163.

71. Zickar MJ, Gibby RE. Four Persistent Themes Throughout the History of I-O Pschology in the United States. In Koppes LL, editor. Historical Perspectives in Industrial and Organizational Psychology. New York: Pschology Press; 2014. p. 61-80.

72. Goldstein H. Assessing group differences. Oxford Review of Education. 1993; 19(2): p. 141-150.

73. Shultz KS, Riggs ML, Kottke JL. The Need for an evolving concept of validity in industrial and personnel psychology: psychometric, legal, and emerging issues. Current Psychology. 1999; 17(4): p. 265-286.

74. Delandshere G. Assessment as inquiry. Teachers College Record. 2002; 104(7): p. 1461-1484.

75. Boud D. Assessment and learning: contradictory or complementary? In Knight P, editor. Assessment for Learning in Higher Education. London: Kogan Publications; 1995. p. 35-48.

76. Stobart G. Validity in Formative Assessment. In Gardner J, editor. Assessment and Learning. 2nd ed. London: Sage Publications; 2012. p. 234-243.

77. Govaerts MJB, Van der Vleuten CPM, Schuwirth LWT, Muijtsens AMM. Broadening perspectives on clinical performance assessment: Rethinking the nature of intraining assessment. Advances in Health Sciences Education. 2007; 12: p. 239-260.

78. Schuwirth L, Van der Vleuten C. A plea for new psychometric models in educational assessment. Medical Education. 2006; 40: p. 296-300.

79. Yorke M. Summative assessment: dealing with the 'measurement fallacy'. Studies in Higher Education. 2011; 36(3).

80. Dochy F. The Edumetric Quality of New Modes of Assessment: Some Issues and Prospects. In Joughin G. Assessment, Learning and Judgement in Higher Education: Springer; 2009. p. 85-114.

81. Dierick S, Dochy F. New lines in edumetrics: new forms of assessment lead to new assessment criteria. Studies in Educational Evaluation. 2001; 27: p. 307-329.

82. Baartman LK. 'Assessing the assessment' : Development and use of quality criteria for Competence Assessment Programmes. Universiteit Utrecht, Utrecht, 2008.

83. Dochy FDJC, McDowell L. Assessment as a tool for learning. Studies in Educational Evaluation. 1997; 23(4): p. 279-298.

84. Torrance H. Postmodernism and educational assessment. In Filer A, editor. Assessment: Social Practice and Social Product. London, GBR: Routledge; 2000. p. 173-188.

85. Hanson AF. How Test Create What They are Intended to Measure. In Filer A, editor. Assessment: Social Practice and Social Product. London, GBR: Routledge; 2000. p. 67-81.

86. James M. Assessment and Learning. In Swaffield S, editor. Unlocking Assessment: Understanding for reflection and application. London: Routledge; 2008.

87. Knight PT. The Achilles' Heel of Quality: The assessment of student learning. Higher Education. 2002b; 8(1): p. 107-115.

88. Shepard LA. Psychometricians' belief about learning. Educational Researcher. 1991; 20(6): p. 2-16.

89. William D. Toward a philosophy for educational assessment. In British Educational Research Association's 20th Annual Conference; 1994; Oxford.

90. Van der Vleuten CPM, Schuwirth LWT, Scheele F, Driessen EW, Hodges B. The assessment of professional competence: building blocks for theory development. Best Practice & Research Clinical Obstetrics and Gynaecology. 2010;: p. 1-17.

91. Borsboom D, Mellenbergh GJ. Test Validity in Cognitive Assessment. In Leighton JP, Gierl MJ, editors. Cognitive Diagnostic Assessment for Education. Cambridge, UK: Cambridge University Press; 2007. p. 85-115.

92. Broadfoot P, Black P. Redefining assessment? The first ten years of assessment in education. Assessment in Education. 2004; 11(1): p. 7-26.

93. Moss PA. Shifting conceptions of validity in educational measurement: Implications for Performance. Educational Research. 1992; 62(3): p. 229-258.

94. Sambell K, McDowell L, Brown S. "But is it fair?": An exploratory study of student perceptions of the consequential validity of assessment. Studies in Educational Evaluation. 1997; 23(4): p. 349-371.

95. Dochy F, Segers M, Gijbels D, Struyven K. Assessment engineering: Breaking down barriers between teaching and learning, and assessment. In Boud F, Falchikow N. Rethinking

Assessment in Higher Education. NewYork: Routledge; 2007.

96. Pellegrino JW. Assessment as a positive influence on 21st century teaching and learning: A systems approach to progress. Psicología Educativa. 2014;(20): p. 65-77.

97. Ecclestone K, Pryor J. 'Learning careers' or 'assessment careers'? The impact of assessment systems on learning. Assessment in Education. 2001; 29(4): p. 471-488.

98. Markus KA, Borsboom D. Frontiers of Test Validity Theory Measurement, Causation, and Meaning New York: Routledge; 2013.

99. Camara WJ, Lane S. A historical perspective and current views on the Standards for Educational and Psychological Testing. Educational Measurement: Issues and Practice. 2006; 25(3): p. 35-41.

100. Plake BS, Wise LL. What is the role and importance of the revised AERA, APA, NCME Standards for. Educational Measurement: Issues and Practice. 2014; 33(4): p. 4-12.

101. Newton PE, Shaw SD. Standards for talking and thinking about validity. Psychological Methods. 2013; 18(3): p. 301-319.

102. Lissitz RW, Samuelsen K. A suggested change in terminology and emphasis regarding validity and education. Educational Researcher. 2007; 38(8): p. 437-448.

103. Moss PA. Reconstructing Validity. Educational Researcher. 2007; 36(8): p. 470-476.

104. Kane MT. Validating interpretations and uses of test scores. Journal of Educational Measurement. 2013; 50(1): p. 1-73.

105. Kane M. An Argument-based Approach to Validation. ACT Research Report Series. Iowa City: The American College Testing Program; 1990.

106. Shay S. Beyond social constructivist perspectives on assessment: the centring of. Teaching in Higher Education. 2008; 13(5): p. 595-605.

107. Brennan RL. Commentary on "Validating the interpretations and uses of test scores". Journal of Educational Measurement. 2013; 50(1): p. 74-83.

108. Sireci S, Padilla JL. Validating assessments: introduction to the special section. Psicothema. 2014; 26(1): p. 97-99.

109. Messick S. Validty. Princeton, New Jersey: Educational Testing Service; 1987.

110. Messick S. Validity of Psychological Assessment: Validation ofInferences from Persons' Responses and Performancesas Scientific Inquiry into Score Meaning. Princeton, N.J.: Educational Testing Service; 1994.

111. Messick S. Meaning and values in test validation: The science and ethics of assessment. Princeton, NJ: Educational Testing Service; 1988.

112. Shepard LA. The Centrality of test use and consequences for test validity. Educational Measurement: Issues and Practice. 1997; 16(2): p. 5-24.

113. Borsboom D, Cramer AOJ, Kievit RA, Scholten AZ, Franic S. The End of Construct Validity. In Lissitz RW, editor. The Concept of Validity: Revisions, New Directions, and Applications. Charlotte, NC: Information Age Publishing; 2009. p. 135-170.

114. Popham JW. Consequential validity: right concern-wrong concept. Educational Measurement: Issues and Practice. 1997; 16(2):

p. 9-13.

115. Kane M. Validating score interpretations and uses. Language Testing. 2012; 29(1): p. 3-17.

116. Shaw S, Crisp V. Reflections on a framework for validation- Five years on. Research Matters: A Cambridge Assessment Publication. 2015;(19): p. 31-37.

117. Kane M. Terminology, Emphasis, and utility in validation. Educational Researcher. 2007; 37(2): p. 76-82.

118. Nichols PD, Meyers JL, Burling KS. A framework for evaluating and planning assessments intended to improve student achievement. Educational Measurement: Issues and Practice. 2009; 28(3): p. 14-23.

119. Perie M, Marion S, Gong B. Moving toward a comprehensive assessment system: A framework for considering interim assessments. Educational Measurement: Issues and Practice. 2009; 28(3): p. 5-13.

120. Fook J, White S, Gardner F. Critical reflection: a review of contemporary literature and understandings. In Gardner F, Fook J, White S. Critical Reflection in Health and Social Care. Berkshire: McGraw-Hill Education Open University Press; 2006. p. 3-20.

121. Mezirow J. How Critical Reflection Triggers Transformative Learning. In Mezirow J, editor. Fostering Critical Reflection in Adulthood. San Fransisco: Josey-Bass p. 1-20.

122. Fook J, White S, Gardner F. Critical reflection: a review of contemporary literature and understandings. In Gardner F, Fook J, White S. Critical Reflection in Health and Social Care. Berkshire: McGraw-Hill Education Open University Press; 2006.

123. Sandars J. The use of reflection in medical education: AMEE Guide No. 44. Medical Teacher. 2009; 31; 31: p. 685–695.

124. Aronson L. Twelve tips for teaching reflection at all levels of medical education. Medical Teacher. 2011;(33): p. 200–205.

125. Henderson E, Berlin A, Freeman G, Fuller J. Twelve tips for promoting significant event analysis to enhance reflection in undergraduate medical students*. Medical Teacher. 2002; 24(2): p. 121–124.

126. Mann K, Gordon J, MacLeod A. Reflection and reflective practice in health professions education: a systematic review. Advances in Health Sciences Education. 2009;(14): p. 595–621.

127. Maudsley G, Strivens J. Promoting professional knowledge, experiential learning and critical thinking for medical students. Medical Education. 2000; 34: p. 535-544.

128. Wald HS, Davis SW, Reis SP, Monroe AD, Borkan JM. Reflecting on reflections: Enhancement of medical education curriculum with structured field notes and guided feedback. Academic Medicine. 2009; 84(7): p. 830-837.