ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2017; 8(2);183-199

Türkiye'de Eğitim Alanında Yayımlanan Ölçek Geliştirme Çalışmalarının Uygunluğunun Çok Yüzeyli Rasch Modeli ile İncelenmesi*

Investigation of Scale Development Studies Conducted in Educational Sciences Published in Turkey by Many-Faceted Rasch Model

Gülden KAYA UYANIK** Neşe GÜLER*** Gülşen TAŞDELEN TEKER****
Süleyman DEMİR*****

Öz

Bu çalışmanın amacı Türkiye'de 2010-2015 yılları arasında sosyal bilimler atıf indeksinde (Social Sciences Citation Index- SSCI) taranan bilimsel dergilerde yayımlanmış ölçek geliştirme çalışmalarının, ölçek geliştirme sürecindeki adımları ne derece sağlayabildiklerini incelemektir. Araştırma kapsamında Türkiye'de eğitim alanındaki üç dergide 2010-2015 yılları arasında yayınlanan 57 ölçek geliştirme makalesi, Ölçme ve Değerlendirme alanında dört uzman tarafından araştırmacıların geliştirdiği "Ölçek Geliştirme Makaleleri İnceleme Formu" ile puanlanmıştır. Ölçek Geliştirme Makaleleri İnceleme Formu alan yazında olan genel görüşler dikkate alınarak 19 maddeden oluşan bir form olarak düzenlenmiştir. Analizler çok yüzeyli Rasch modeli (ÇYRM) için FACETS ve frekans analizi için SPSS programı kullanılarak gerçekleştirilmiştir. Çalışmada yer alan üç yüzeyin (makalelerin belirlenen ölçütleri karşılama düzeyi, değerlendirme formunda kullanılan ölçütler ve puanlayıcılar) ve değerlendirmedeki ölçütlerin puanlama düzeylerine ait bulgular ayrı ayrı elde edilmiştir. Analizler sonucunda, Logit cetvele göre çalışmada esas alınan ölçütleri karşılayabilme düzeyi en yüksek 14 numaralı makalenin ve en düşük 27 numaralı makalenin olduğu gözlenmektedir. Değerlendirme formunda yer alan ölçütler incelendiğinde güçlük düzeyi en yüksek (karşılanması en zor) 16. ölçüt (*Ölçüt geçerliği analizi sonucu elde edilen değerler uygundur)* iken güçlük düzeyi en düşük (karşılanması en kolay) 3. ölçüt (*Ölçeğin amacına uygun bir hedef kitle belirlenmiştir.*) olduğu belirlenmiştir. Puanlayıcılara ait ölçümler incelendiğinde ise 4. puanlayıcının en katı (en yüksek pozitif değer), 1. puanlayıcının ise en cömert (en düşük negatif değer) puanlayıcı olduğu gözlenmektedir. Ayrıca incelenen 57 makalenin 27 tanesinde (%47,4) geliştirilen ölçekler makale içeriğinde verilirken 30 makalede (%52,6) bu durum gözlenmemiştir. Geliştirilen ölçeklerin diğer araştırmacılar tarafından kullanıldığında puanlamanın nasıl yapılacağına dair bilginin verilme durumu ise makalelerin %56,1'inde (32 makale) gözlenirken % 43,9'unda (25 makale) puanlama ile ilgili bir bilgiye rastlanmamıştır.

Anahtar Kelimeler: Ölçek geliştirme, çok yüzeyli rasch modeli, eğitim bilimleri

Abstract

The aim of this study is to investigate the scale development studies that were conducted in Turkey from 2010 to 2015 and published in the journals indexed in Social Sciences Citation Index (SSCI) by means of their verifications about scale development steps. In the study, 57 scale development articles published in between years 2010-2015 in three different journals scope of whom are educational sciences were rated according to "Scale Development Article Investigation Form" by four measurement and evaluation expert. The 19-item Scale Development Article Investigation Form was composed according to the view of the literature. The many-facet Rasch model analysis were done by using FACETS and frequencies were gathered by using SPSS programs. The analysis results of the three facets (the level of articles by means of verification of determined

* Çalışmanın bir kısmı Eğitimde ve Psikolojide Ölçme Ve Değerlendirme (Antalya / 2016) kongresinde sunulmuştur.

**Yrd. Doç. Dr., Sakarya Üniversitesi, Sakarya, Türkiye, guldenk@sakarya.edu.tr
***Doç. Dr., Sakarya Üniversitesi, Sakarya, Türkiye, gnguler@gmail.com.tr
****Yrd. Doç. Dr. Hacettepe Üniversitesi, Ankara, Türkiye, gulsentasdelen@gmail.com
*****Araş. Gör., Sakarya Üniversitesi, Sakarya, Türkiye, suleymand@sakarya.edu.tr

DOI: 10.21031/epod.291367
Geliş Tarihi: 11.02.2017
Kabul Tarihi: 13.03.2017

criteria, criteria used in investigation form, and the raters) used in the study and the rating levels of investigated criteria were gathered separately. As a result of the study it was seen that the 14th article was cover the investigated criteria mostly and the 27th article was the least cover one according to the logit scale. When the criteria used for investigation of the articles were examined, the 16th criteria was the most difficult one. On the other hand, the third criteria (There were a suitable sample for the aim of scale.) was the easiest one to meet. When the ratings of raters were investigated, it was observed that the fourth rater was the strictest one and the first rater was more lenient one. Moreover, the developed scales were given in just 27 of 57 investigated articles (47,4%) and there were no such a situation on 30 articles (52,6%). While the scoring information of developed scales for other researchers who want to use the developed scales in their future studies was given only 56,1% (f=32) of investigated articles, there were no such information for 43,9 % of investigated articles.

Keywords: Scale development, many facet Rasch model, educational sciences

GİRİŞ

Sosyal bilimlerdeki alan çalışmalarında, bireylerin farklı yetenek, algı ve tutumlarını ölçmek, bunlara dayanarak bazı kararlar almak ve farklı yapılar arasındaki ilişkileri açıklamak amacıyla veri toplama aracı olarak sıklıkla ölçek adı verilen araçlardan yararlanılmaktadır (Stone, 1978). Ölçme aracı olarak ölçeklerin kullanılmasının temeldeki sebebi, eğitimde ve psikolojide var olduğu kabul edilen ancak doğrudan gözlenemeyen değişkenlerin ölçülmek istenmesidir. Bireyde var olduğu kabul edilen bu tür değişkenler yapı olarak adlandırılmakta ve bu yapılar bireylerde gözle görülür davranışlara neden olmaktadırlar. Ölçekler ile gözle görülür davranışların dereceli puanlama anahtarları kullanılarak bireylerin ilgili yapıya ne kadar sahip olduklarını belirlemek amaçlanmaktadırlar (DeVellis, 2003; Erkuş, 2012; Tezbaşaran, 2008).

Ölçeklerin bireylere uygulanması sonucu elde edilen bulguların genellenebilirliği, işlevselliği ve sağlamlığı ise kullanılan bu araçların güvenirliği ve geçerliğiyle paralellik göstermektedir (Erkuş, 2007). Örneğin aynı ölçeklerin veya aynı özelliği ölçmek üzere geliştirilen farklı ölçeklerin kullanıldığı çalışmalarda elde edilen sonuçlar birbirinden tutarsız olabilmektedir. Bu durumun temelde en büyük sebebinin güvenilir ve geçerli ölçümler sağlayan ölçme araçlarının kullanılmamış olmasından kaynaklandığı düşünülmektedir (Barrett, 1972; Cook, Hepworth, Wall ve Warr, 1981; Hinkin, 1995; Schriesheim, Powers, Scandura, Gardiner ve Lankau, 1993). Bu sebeple doğrudan gözlenemeyen değişkenleri ölçmek için kullanılan bu araçların niteliği oldukça önemlidir.

Araştırmalarda kullanılan ölçekler belirli bir özelliği ölçmeye yönelik olarak geliştirilebileceği gibi daha önceden araştırmacının ilgilendiği yapıyı ölçmek amacıyla başka bir dilde ve kültürde geliştirilmiş ölçekler de araştırmacının ilgilendiği örnekleme uygun olarak uyarlanarak kullanılabilir. Bu süreçlerden ilki ölçek geliştirme olarak tanımlanırken ikincisi ölçek uyarlama olarak ifade edilmektedir. Araştırmacılar bu iki süreçten araştırmalarında ölçmeyi düşündükleri değişkene göre uygun olanı seçerek çalışmalarını yürütmelidirler. Eğer ölçmeyi planladıkları değişkene ilişkin daha önce geliştirilmiş herhangi bir ölçek bulunmuyorsa ya da daha önce geliştirilmiş olan ölçekler, araştırmacının amacına tam anlamıyla hizmet etmiyorsa yeni bir ölçek geliştirmeleri gerekmektedir. Ancak eğer daha önce başka bir dilde, kültürde veya araştırmacının uygulamak istediği gruptan farklı bir gruba yönelik olarak geliştirilmiş bir ölçek bulunuyorsa, bu ölçeğin araştırmacının uygulamak istediği dile, kültüre ve örnekleme uygun hale getirilmesi gerekmektedir. Bir diğer ifadeyle uyarlanması gerekmektedir. Bu iki sürecin de birbirlerine göre avantaj ve dezavantajları bulunmaktadır. Bu çalışma kapsamında ölçek geliştirme makaleleri incelendiğinden ölçek uyarlama üzerinde fazla durulmamış ölçek geliştirme ise daha detaylı olarak irdelenmiştir.

Araştırmacının amacına uygun bir ölçek bulunmadığı durumda alanyazında belirtilen ölçek geliştirme adımlarını takip ederek yeni bir ölçek oluşturulmaya çalışılır. Bir diğer ifadeyle güvenilir ve geçerli ölçme araçlarının geliştirilmesi için dikkat edilmesi ve eksiksiz bir şekilde izlenmesi gereken bazı işlem basamakları bulunmaktadır. Bu aşamalar alanyazında aşağıdaki gibi özetlenmiştir (Clark ve Watson, 1995; DeVellis, 2003; Hinkin, 1995; Hinkin, Tracey ve Enz, 1997; Tezbaşaran, 2008):

- Ölçme aracına ilişkin açık ve net bir yönergenin hazırlanması,

- Ölçülmek istenen özelliğin açık ve net olarak ifade edilmesi,
- Madde havuzunun oluşturulması,
- Madde havuzunun uzman görüşüne sunulması,
- Ön deneme uygulamasının yapılması,
- Deneme uygulamasının gerçekleştirilmesi,
- Faktör yapısının belirlenmesi,
- Güvenirliğinin belirlenmesi
- Ortaya atılan modelin doğrulanması

Özellikle son yıllarda ölçek geliştirme ve uyarlama çalışmalarında bir artış gözlenmektedir. Erkuş (2007) bu artışın sebebi olarak akademisyenlerin üzerindeki yayın baskısını göstermektedir. Bu durum ise niteliği düşük çalışmalara yol açmaktadır. Çünkü yürütülen çalışmalar incelendiğinde süreçte yapılan ve yapılmaya devam edilen bir takım sorunların ve eksiklerin olduğu gözlenmektedir. Bu sorunları ve eksiklikleri belirleyip ölçek geliştirmek veya uyarlama yapmak isteyen araştırmacılara ışık tutmak adına alanyazında yürütülmüş bazı çalışmalar bulunmaktadır. Bu çalışmaların bir kısmı hem ölçek geliştirme hem de uyarlama çalışmalarını incelerken (Acar Güvendir ve Özer Özkan, 2015; Çüm ve Koç, 2013; Erkuş, 2007) sadece uyarlama (Boztunç Öztürk, Eroğlu ve Kelecioğlu, 2014) ya da sadece ölçek geliştirme çalışmalarına odaklanan (Tavşancıl, Güler ve Ayan, 2014) çalışmalarda bulunmaktadır. Bu çalışma kapsamında özellikle ölçek geliştirmeye odaklanan çalışmalar incelenmiştir.

Çüm ve Koç (2013) 2005-2013 yılları arasında yayımlanan 29 ölçek geliştirme makalesini, ölçek geliştirme adımları açısından incelemişlerdir. Araştırma sonucunda, çalışmaların %67'sinde ölçek geliştirme adımlarına uygun bilgilerin rapor edildiği belirtilmiştir. Erkuş (2007) çalışmasında, ölçek geliştirme ve uyarlama çalışmalarındaki genel sorunlar üzerinde durmuştur. Çalışmanın amacını ise sadece bu konuda bir farkındalık yaratıp bu yönde bir takım girişimlerin başlatılabilmesi olarak ifade etmiştir. Çalışma kapsamında çalışmanın başlığı, ölçülmek istenen örtük değişkenin kavramsal tanımı, madde üretilmesi, deneme uygulaması, madde analizi, güvenirlik, geçerlik, puanların yorumlanması ve çalışmanın raporlaştırılması başlıkları ile ilgili sorunlar belirtilip olası nedenler ve çözüm önerilerinden de bahsedilmiştir.

Tavşancıl, Güler ve Ayan (2014) eğitim ve psikoloji alanlarında yürütülen çalışmalardaki tutum ölçeklerinin geliştirilme adımlarının karşılanma durumlarını ortaya koymaya çalışmışlardır. Bu amaç doğrultusunda 54 tutum ölçeği belirlenen ölçütlere göre incelenmiş ve alanyazında yer alan çok sayıda yanlış ve eksik tutum ölçeğinin olduğu sonucuna ulaşılmıştır. Bunun yanı sıra, yürütülen bu çalışma ile aynı konuda geliştirilmiş birden fazla ölçeğin tespiti de araştırmacılar için emek ve zaman kaybı olarak belirtilmiştir.

Acar Güvendir ve Özer Özkan (2015) ise 2006-2014 yılları arasında "Social Science Citation Index" (SSCI) dizininde taranan Türkiye'de eğitim alanında yayımlanan üç dergideki 26 ölçek geliştirme makalesini süreçte izlenen adımlara göre incelemişlerdir. Elde edilen bulgulara göre çalışmaların tümünde iç tutarlılık güvenirlik belirleme yöntemlerinden Cronbach alfa güvenirlik katsayısının kullanıldığı tespit edilmiştir. Ayrıca, incelenen çalışmaların çok azında geliştirilen ölçeğin uygulama yönergesinin bulunduğu belirlenmiştir.

Bu çalışmada da Türkiye'de eğitim alanında SSCI dizininde yer alan üç dergide 2010-2015 yılları arasında yayınlanan ölçek geliştirme makaleleri, "ölçek geliştirme süreci"ndeki adımları ne derece sağladıkları açısından, dört ölçme ve değerlendirme alan uzmanı tarafından incelenmiştir. Uzmanlardan elde edilen ölçme sonuçlarının güvenirliğinin belirlenmesinde çok yüzeyli Rasch modelinden yararlanılmıştır.

Çok Yüzeyli Rasch Modeli

Madde Tepki Kuramı (MTK)'na dayalı modellerden biri olan Rasch Modeli bir parametreli (iki yüzeyli) bir modeldir. Ölçme sonuçlarının, birey ve madde yüzeyleri dışında farklı değişkenlik kaynakları (puanlayıcılar, farklı ölçme oturumları vb.) tarafından etkilendiği ölçme durumlarında ise "Çok Yüzeyli Rasch Modeli"nden yararlanılır. Çok yüzeyli Rasch Modeli (ÇYRM)'de yer alan her bir değişkenlik kaynağına "yüzey" adı verilir. Bu sebepledir ki birden fazla değişkenlik kaynağının yer aldığı ölçme durumlarında ele alınan bu model "Çok Yüzeyli Rasch Modeli" olarak ifade edilir. ÇYRM, Klasik Test Kuramı (KTK)'nın sınırlı kaldığı noktaların üstesinden gelebilen bir ölçme modelidir (Anshel ve diğerleri., 2009; Baştürk ve Işıkoğlu, 2008; Güler, 2014; Güler ve Gelbal, 2010; Kim, Park ve Kang, 2012). MTK'ya dayalı tüm modellerde olduğu gibi ÇYRM'de de her bir yüzeye (birey, madde, puanlayıcı, durum vb.) ilişkin kestirimler, diğer değişkenlik kaynaklarından bağımsız yapılabilmektedir. Örneğin; bireylere ilişkin elde edilen ölçümler, ölçme aracındaki maddelerin güçlük düzeylerinden, puanlayıcıların katılık/cömertliklerinden vb. değişkenlik kaynaklarından; madde parametreleri uygulandığı bireylerin ölçülen özelliğinden, puanlayıcıların katılık/cömertlik düzeylerinden ve ölçme sonuçlarını etkileyebilecek diğer değişkenlik kaynaklarından bağımsız olarak kestirilebilir (Engelhard, 1994). KTK'da bireylerin bir testteki yetenek düzeyleri, test maddelerinden aldıkları puanların toplamı ile kestirilir. Testteki her bir maddenin güçlük düzeylerinin (ya da bir maddeye katılma ya da katılmama olasılıklarının) birbirine eşit ve/veya toplam puana olan katkısının aynı olduğu varsayılır. Bununla birlikte her madde, ölçülen özellikte farklı bir katkıya sahipse her bir maddenin katkısını, toplam puanda eşit kabul etmek yanlı sonuçlara sebep olur ve bu kabul üzerinden yapılan istatistikler şüphe içerir (Anshel ve diğerleri, 2009, Brinthaupt ve Kang, 2012). KTK'da ham puanlardan elde edilen sonuçlar üzerinden bireyler yetenek vb. düzeylerine göre ancak sıralanabilir; sıralama ölçeğinde elde edilen bu puanlar ise toplanabilir değildir. Hâlbuki ÇYRM'nin dayandığı matematiksel model bu sınırlılığın üstesinden gelmeye yardımcı olur. Bu matematiksel model ile ham verilerin doğal logaritması alınır (log-odds) ve böylece ölçme sonuçları, eşit aralıklı ölçek (logit) düzeyine dönüştürülür. KTK ile kayıp verilerin bulunduğu ölçme durumlarında, ham veriler üzerinden yapılan analizler hatalı sonuçlar üretebilirken; ÇYRM'de kayıp veriler göz önünde bulundurulmaz, sadece gözlenen verilere dayalı analizler gerçekleştirilir. Bunlara ek olarak, ÇYRM'de tek bir analiz ile her bir değişkenlik kaynağı için uyum istatistikleri (INFIT ve OUTFIT) belirlenebilir. Uyum istatistikleriyle, model-veri arasındaki uyumun derecesini inceleyebilmek mümkün olmaktadır (Revesz, 2012; Smith ve Kulikowich, 2004). Ayrıca, her bir yüzeye ilişkin parametre kestirimleri ortak bir cetvel (logit cetvel) üzerinde birlikte yorumlanabilir (Linacre, 1989). Ortak bir metriğe sahip bu cetvelde yüzeylerin göreli yerleri incelenebilir. Böylece, örneğin; maddelerin dağılımı gözlenerek, yetenek düzeyi boyunca hangi düzeyde madde yok/eksik kalmış, hangi düzeyde çok sayıda madde yer almış belirlenebilir (Brinthaupt ve Kang, 2012). Bunun yanı sıra, çalışmada yer alan diğer yüzeylerle (örneğin; puanlayıcılarla) ilgili de açıklayıcı bilgiler sunar. Örneğin, birden fazla puanlayıcının yer aldığı bir ölçme durumunda, bir puanlayıcı diğerlerinden daha cömert puanlama yaparken; diğer tüm puanlayıcıların yüksek puan verip, bu puanlayıcının daha düşük puan verdiği bu "beklenmedik puanlama durumu" tespit edilebilir. Ayrıca, ÇYRM ile puanlayıcıların katılık/cömertlik düzeylerine ilişkin bilgi almak mümkün olduğu gibi halo etkisi, merkeze yönelme ve/veya yanlılık gibi diğer puanlayıcılardan kaynaklanan hatalar da belirlenebilir (İlhan, 2016). Böylece bu tür incelemelerle ÇYRM; puanlamada alınabilecek önlemlere, puanlayıcılara verilecek eğitime vb. ilişkin aydınlatıcı bilgiler sunar.

ÇYRM'de her bir yüzey için elde edilen diğer istatistikler ise ayırma indeksi ve güvenirlik katsayısıdır. Ayırma indeksi, 1 ile ∞ arasında değer alabilirken güvenirlik katsayısı KTK'da da olduğu gibi 0 ile 1 arasında değişmektedir (Sudweeks, Reeveb ve Bradshaw, 2005). Birey ve madde yüzeyi için ayırma indeksi ve güvenirlik katsayısının yüksek değerlere sahip olması istenilen bir durum olup puanlayıcı yüzeyi için bu değerlerin düşük olması istenir. Birey yüzeyi için güvenirlik değeri, Cronbach α değerine benzer yorumlanmaktadır ve her iki istatistiğin yüksek değerlere sahip olması bireylerin birbirinden ayırt edilebildiğine işaret eder. Madde yüzeyi için elde edilen ayırma indeksi ve güvenirlik katsayısının yüksek çıkması farklı özellikleri ölçen maddelerin birbirinden

ayırtedilebildiğini gösterirken; puanlayıcı yüzeyi için hesaplanan bu iki değerin yüksek olması ise puanlayıcıların bireyleri oldukça farklı puanladıklarını ifade eder (İlhan, 2016).

Araştırmanın Amacı

Bu çalışmada, Türkiye'de 2010-2015 yılları arasında SSCI'da yer alan bilimsel dergilerde yayımlanmış ölçek geliştirme çalışmalarının, "ölçek geliştirme süreci"ndeki adımları ne derece sağlayabildiklerine ilişkin bilgi sunmak amaçlanmıştır. Bu amaç için öncelikle, elde edilen verilerin ÇYRM ile her bir yüzey için güvenirliği incelenmiş, sorun teşkil eden durumlar gözden geçirilmiş ve sonrasında elde edilen yeni veriler analiz edilmiştir. Böylece, makaleleri değerlendiren puanlayıcı yüzeyinin, puanlamada kullanılan ölçütlerin ve incelenen makalelerin düzeyleri ve güvenirlikleri belirlenmiştir. Ardından, çalışmada yer alan "ölçek geliştirme" makalelerinin belirlenen ölçütler açısından eksik ve güçlü yönleri tanımlanmaya çalışılmıştır. Alanyazında yer alan, ölçek geliştirme makalelerinin incelendiği çalışmalara destekleyici bilgiler sunmanın yanı sıra yapılan bu çalışmada ÇYRM ile tüm yüzeylerin analiz edilerek; özellikle çalışmada yer alan makalelerin ve belirlenen ölçütlerin eşit aralıklı ölçek düzeyindeki yerleri belirlenmiştir. Elde edilen bu bulgular ile incelenen tüm çalışmalar aldıkları puanlar açısından, ayrıca incelenen ölçütler karşılanma oranları bakımından sıralanmıştır. Çalışmaların ve ölçütlerin cetvel üzerinde yerlerinin belirlenmesinin alan yazına bu ölçek geliştirme çalışmalarının değerlendirilmesi açısından katkı sağlayacağı düşünülmektedir. Buna ek olarak, ÇYRM'nin benzer ölçme durumlarında kullanılabileceğine ilişkin örnek bir çalışma olabilmesi amaçlanmıştır.

YÖNTEM

Bu çalışmada Türkiye'de eğitim konusunu temel alan ve SSCI kapsamında taranan üç derginin (Eğitim ve Bilim, Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, Kuram ve Uygulamada Eğitim Bilimleri Dergisi) 2010-2015 yılları arasında yayınladığı ölçek geliştirme makaleleri incelenmiştir. Araştırmada yayınlanmış çalışmalar var olduğu şekilde tanımlandığı için çalışma tarama modelinde betimsel bir araştırmadır (Karasar, 1998).

Örneklem/Çalışma Grubu

Araştırma kapsamında Türkiye'de eğitim alanında SSCI dizininde yer alan üç dergide 2010-2015 yılları arasında yayınlanan ölçek geliştirme makaleleri çalışmanın evrenini oluşturmaktadır. Bu evrenden amaçlı örnekleme yöntemlerinden biri olan ölçüt örnekleme yolu ile örneklem belirlenmiştir. Ölçüt örnekleme yönteminde bir takım ölçütler önceden belirlenir ve bu ölçütleri karşılayan tüm durumlar çalışmaya katılır (Yıldırım ve Simsek, 2008). İncelenen makalelerin çevrim içi tam metinlerinin ulaşılabilir olması belirlenen ölçütlerden biridir. Çalışmaya, esas amacı ölçek geliştirmek olmayan ancak içinde veri toplama aracı olarak geliştirilmiş ölçek olan çalışmalar, bilgilerin sınırlı olabileceği gerekçesi ile dâhil edilmemiştir. Çalışmaya sadece tipik davranışları ölçmek amacıyla geliştirilen ölçekler ve bu ölçeklerin geliştirilme süreçlerinin rapor edildiği çalışmalar dâhil edilmiştir. Bu bağlamda SSCI kapsamında olan ve eğitim içerikli üç dergide (Eğitim ve Bilim Dergisi, Hacettepe Üniversitesi Eğitim Fakültesi Dergisi ve Kuram ve Uygulamada Eğitim Bilimleri Dergisi) 2010-2015 yılları arasında tam metinlerine ulaşılabilen 57 ölçek geliştirme makalesi, çalışmanın örneklemini oluşturmaktadır. Çalışmaların yıllara göre dağılımı Tablo 1'de verilmiştir.

Tablo 1. Çalışmada kullanılan makalelerin yıllara göre dağılımı

	2010	2011	2012	2013	2014	2015	Toplam
Eğitim ve Bilim Dergisi	4	2	2	3	10	2	23
Hacettepe Üniversitesi Eğitim Fakültesi Dergisi	3	4	2	2	1	2	14
Kuram ve Uygulamada Eğitim Bilimleri Dergisi	-	6	5	7	2	-	20
Toplam	7	12	9	12	13	4	57

Tablo 1'de ölçek geliştirme çalışmaları en çok 2014 yılında yapılmışken, en az çalışmanın 2015 yılında yayınlandığı görülmektedir. Ayrıca çalışmada incelenen 57 makalenin 23'ü Eğitim ve Bilim 14'ü Hacettepe Üniversitesi Eğitim Fakültesi ve 20'si Kuram ve Uygulamada Eğitim Bilimleri dergisinde yayınlanmıştır.

Veri Toplama Araçları

Araştırmada, ölçek geliştirmenin adımları ve gereklilikleri konusunda yayınlanmış kaynakların (Cohen ve Swerdlik, 2010; Edenborough, 1999; Erkuş, 2012; Murphy ve Davidshofer, 2005; Tezbaşaran, 2008) ve ölçek geliştirme makalelerinin incelendiği çalışmaların (Acar Güvendir ve Özer Özkan, 2015; Tavşancıl, Güler ve Ayan, 2014; Çüm ve Koç, 2013; Erkuş, 2007) taraması yapılmıştır. Ölçek geliştirme adımları hakkında alanyazında olan genel görüşler dikkate alınarak araştırmacılar tarafından "Ölçek Geliştirme Makaleleri İnceleme Formu" oluşturulmuştur. Bu formun hazırlanmasında alanyazında bulunan ölçek geliştirme makalelerinin incelendiği çalışmalarda kullanılan ölçütlerden de yararlanılmıştır. Ancak bu ölçütlerden bir kısmının her ölçek geliştirme çalışmasında rapor edilmesinin zorunlu olmadığı düşünüldüğünden daha önceki çalışmalarda esas alınan ölçütlerin bir kısmı bu çalışma kapsamında geliştirilen forma dâhil edilmemiştir. Örneğin, Acar Güvendir ve Özer Özkan'ın (2015) yürüttükleri çalışmada "iki eş değer yarı güvenirliğinin hesaplanmış olması" ölçek geliştirme çalışmalarının incelenmesinde kullanılan ölçütlerden biri olarak belirlenmiştir. Bu çalışmada ise iki yarı güvenirliğinin her ölçek geliştirme çalışmasında rapor edilmesinin zorunlu olmadığı düşünülmüş ve bu sebeple ölçek geliştirme çalışmalarının incelemek üzere geliştirilen formda böyle bir ölçüte yer verilmemiştir. Benzer şekilde, Delice ve Ergene'nin (2015) matematik eğitimi alanındaki ölçek geliştirme çalışmalarını inceledikleri araştırmada, ölçeğin geneline ilişkin Cronbach alfa güvenirlik katsayısının rapor edilmesi gibi bir ölçüt esas alınmaktadır. Ancak tüm ölçeklerde her zaman toplam puan almanın mümkün olmayabileceği ve dolayısıyla ölçeğin geneline ilişkin Cronbach alfa katsayının rapor edilmesinin her ölçek için bir zorunluluk olmadığı düşünüldüğünden bu çalışma kapsamında ölçek geliştirme çalışmalarının incelenmesinde kullanılan forma, bu tür bir ölçüt dâhil edilmemiştir.

Oluşturulan formun kapsamını değerlendirmesi için dört ölçme ve değerlendirme, dilsel uygunluğu açısından değerlendirme yapması için ise iki dilbilgisi uzmanından görüş alınmıştır. Alınan görüşler doğrultusunda form tekrar düzenlenerek kullanıma hazır hale getirilmiştir.

Elde edilen ölçek geliştirme makaleleri inceleme formu 19 maddeden oluşmaktadır ve ilk 17 maddesi 0 - 3 arası puanlanabilir ölçütlerdir. Bu ölçütlerdeki puanlama, "(0) bilgi yok; (1) bilgi var ancak hiç açıklanmamış; (2) bilgi var ve kısmen açıklanmış; (3) bilgi var ve gerekli şekilde açıklanmış" olarak tanımlanmıştır. Diğer iki madde de ise geliştirilen ölçeğin çalışmada yer alıp almadığı ve ölçeğin puanlanması ile ilgili bilginin verilip verilmediği sorularak; bu maddeler evet-hayır şeklinde cevaplanabilir biçimde oluşturulmuştur.

Verilerin Analizi

Çalışmada Çok Yüzeyli Rasch Analizi kullanılmıştır. ÇYRM kavramsal olarak regresyon analizine benzerdir. Bağımlı değişken, bireyin (ya da ölçmeye konu olan objenin) bir maddeden alabileceği

puanların olasılıkları oranının lojistik dönüşümü, bağımsız değişkenler ise ölçmede yer alan madde güçlük düzeyi, puanlayıcı katılık/cömertlik düzeyi gibi diğer yüzeylerdir (Randall ve Engelhard, 2009). Böylece, bu çalışmadaki gibi makale, ölçütler ve puanlayıcı gibi üç yüzeyin yer aldığı bir ÇYRM Eşitlik 1'deki gibi ifade edilebilir:

$$\text{Log}\,(P_{mkpc} / P_{mkpc-1}) = B_m - D_k - C_p - F_c \quad (1)$$

P_{mkpc} : m. makalenin k. ölçütte p. puanlayıcı tarafından c kategorisinde puan alma olasılığı

P_{mkpc-1}: m. makalenin k. ölçütte p. puanlayıcı tarafından c-1 kategorisinde puan alma olasılığı

B_m: m. makalenin esas alınan ölçütü karşılama düzeyi

D_k: k. ölçütün güçlük düzeyi

C_p: p. puanlayıcının katılık/cömertlik düzeyi

F_c: c. kategoriden c-1. kategoriye geçişteki güçlük düzeyi

ÇYRM'de bireylerin maddeleri cevaplama olasılıkları (bu çalışmada, makalelerin esas alınan ölçütleri karşılama olasılıkları), yukarıda da değinildiği gibi, "log-odds" olarak ifade edilir ve lojit cetvelde "log-odds" ya da "logit" birimlerle gösterilir. Lojit cetvelde artan pozitif değerler bireyler için yüksek yetenek düzeyine (bu çalışma için makalelerin, belirlenen ölçütleri karşılama düzeylerinin yüksek oluşu), maddeler için yüksek güçlük düzeyine ve puanlayıcılar için yüksek katı puanlama düzeyine karşılık gelir. Logit cetvelin, her bir yüzeyi görsel olarak inceleme olanağı sağlıyor olması; her yüzeydeki elemanların sırasının, yüzeylerin birbirleriyle olan göreli durumlarını ve bir yüzeydeki elemanlar arası farklılıkların ne büyüklükte olduğunu gözleme imkânı sunar (Güler, 2014; Hetherman, 2004).

Çalışmada yer alan 57 makalenin tümü dört ölçme ve değerlendirme uzmanı tarafından ölçek geliştirme makaleleri inceleme formu kullanılarak puanlanmıştır. Formda yer alan ilk 17 değerlendirme düzeyinin analizi, Çok Yüzeyli Rasch Modeli (ÇYRM) ile Linacre (2007) tarafından geliştiren FACETS programı kullanılarak gerçekleştirilmiştir. Analiz sonucunda, çalışmada yer alan üç yüzeyin ("makalelerin belirlenen ölçütleri karşılama düzeyi", "değerlendirme formunda kullanılan ölçütler" ve "puanlayıcılar") ve değerlendirmedeki ölçütlerin puanlama düzeylerinin birlikte yer aldığı logit cetvel sunulmuştur. Ardından tüm yüzeylere ilişkin ölçüm raporları verilmiştir. Ölçüm raporunda yer alan tüm yüzeylere dair uyum-içi ve uyum-dışı istatistiklerinin istenilen değer aralığı olan 0.5 ile 1.5 arasında olması (Turner, 2003; Wright ve Linacre, 1992), ölçüm sonuçlarının tek boyutlu bir yapı gösterdiğine işaret etmektedir (Anshel ve diğerleri; 2009; Lee ve diğerleri, 2010). Uyum istatistiklerine ek olarak, her bir yüzey için elde edilen güvenirlik değerleri ve ayırma oranları da yorumlanmıştır.

ÇYRM'nin FACET programı ile analizinde "beklenmedik tepkilerin (unexpected responses)" incelenmesini sağlayan bir tablo da yer alır. Bu tabloda var olan bilgiler, güvenirliği olumsuz etkileyen beklenmeyen puanların incelenebilmesine imkân tanır (Güler, 2014). Örneğin, dört puanlayıcının yer aldığı bir ölçme çalışmasında, üç puanlayıcı bir makalenin incelendiği formdaki bir maddesine çok yüksek puan verirken diğer puanlayıcı aynı makalenin aynı maddesine en düşük puanı vermişse, bu durumu gözlemlemek bu tablo ile mümkün olmaktadır. Bu tabloda ayrıca standartlaştırılmış artık değerler de rapor edilir. Bu değerler, model ile verinin uyumlu olup olmadığı hakkında karar vermemize yardımcı olur (İlhan, 2016). Alanyazında, model-veri uyumunun olabilmesi için ± 2 sınırları dışında kalan standartlaştırılmış artık değerler sayısının toplam veri sayısının yaklaşık %5'ini ve ± 3 sınırları dışında kalan standartlaştırılmış artık değerleri sayısının ise toplam veri sayısının yaklaşık %1'ini geçmemesi gerektiği belirtilmektedir (Linacre, 2014). Bu çalışmada, yer alan toplam veri sayısı 3876 olup (57 ölçek x 17 madde x 4 puanlayıcı = 3876); ± 3 aralığı dışında kalan standartlaştırılmış artık veri sayısı (31) oranı %0.8'dir. Bu sonuca göre, çalışmada yer alan verilerin model ile uyumlu olduğu söylenebilir. Ancak, ölçekler ile ilgili yapılacak çıkarımların en güvenilir sonuçlar üzerinden yapılabilmesi için analiz sonucu elde edilen 31 beklenmedik veri tekrar incelenmiş, problemli puanlamaların görüldüğü makaleler tekrar puanlayıcılara bildirilmiş ve yeniden puanlamaları istenmiştir. Yeni elde edilen veriler üzerinden

yapılan analizle de (alanyazında da yer aldığı üzere) beklenmedik verilerin çok az olmasından dolayı, tüm bulguların aynı değerlere sahip olduğu, sadece makalelere ilişkin elde edilen ölçüm raporundaki güvenirlik değerinin 0.85'ten 0.88'e yükseldiği gözlenmiştir.

Bu analizlerin ardından, çalışmanın kapsamında yer alan 57 ölçeğe ilişkin elde edilen logit puanlar, yöntem kısmında detaylı olarak açıklandığı şekilde değerlendirme düzeyleri olan 0-3 değerlerine dönüştürülerek yorumlanmıştır. Benzer süreç, ayrıca değerlendirme formunda yer alan ölçütler için de gerçekleştirilmiş ve her ölçüt [0, 1] puan aralığında karşılık gelen değeri üzerinden incelenmiştir. Bu işlemler için Excel bilgisayar programından yararlanılmıştır.

Son olarak çalışmada kullanılan makale inceleme formunda yer alan "Geliştirilen ölçek çalışmada yer almış." ve "Geliştirilen ölçeğin nasıl puanlanacağı hakkında bilgi verilmiş." ifadelerine puanlayıcıların verdikleri *Evet/Hayır* cevaplarına ilişkin verilere SPSS 20 programı kullanılarak frekans ve yüzde hesaplanmıştır.

BULGULAR

Verilerin ÇYRM ile Analizi

Bu bölümde araştırma sonucu elde edilen bulgulara yer verilmiştir. Öncelikle, FACETS programı ile gerçekleştirilen ÇYRM analizi sonuçları verilmiştir. Şekil 1'de, düzeltilmiş verilere ilişkin elde edilen logit-cetvel sunulmuştur.

```
+-------------------------------------------------------------------------+
|Measr|+MAKALE                          |-MADDE      |-PUANLAYICI|Scale|
|-----+---------------------------------+------------+-----------+-----|
|    2 +                                 +            +           + (3) |
|      |                                 |            |           |     |
|      |                                 |            |           |     |
|      |                                 |            |           |     |
|      |                                 |            |           |     |
|      |                                 |            |           |     |
|      |                                 |            |           |     |
|      |                                 |            |           |     |
|      | 14                              |            |           |     |
|    1 +                                 +            +           +     |
|      |                                 |            |           |     |
|      | 3                               |            |           |     |
|      | 16 36                           | 16         |           | ---   |
|      | 38 44                           |            |           |     |
|      |                                 | 14         |           |     |
|      | 53                              | 5          | 4         |     |
|      | 24 25 28 40 43                  | 7          |           |  2  |
|      | 5  12 39 50 57                  |            |           |     |
|      | 1  8  10 13 29 37 41 46 47 51   |            | 3         |     |
*    0 * 6  15 21 22 42 45 49 55 56    *            * 2         * --- *
|      | 7  19 30 32 52                  | 9          |           |     |
|      | 9  18 20 48                     |            |           |     |
|      | 17 31 34 35                     | 15         |           |  1  |
|      |                                 |            |           |     |
|      | 54                              |            | 1         |     |
|      | 2  11 23 26                     |            |           |     |
|      |                                 | 8          |           | ---   |
|      | 4                               | 11         |           |     |
|      | 33                              | 6          |           |     |
|   -1 + 27                              + 4          +           +     |
|      |                                 | 2          |           |     |
|      |                                 | 17         |           |     |
|      |                                 | 13         |           |     |
|      |                                 |            |           |     |
|      |                                 | 1   10  12 |           |     |
|      |                                 | 3          |           |     |
|      |                                 |            |           |     |
|      |                                 |            |           |     |
|      |                                 |            |           |     |
|   -2 +                                 +            +           + (0) |
|-----+---------------------------------+------------+-----------+-----|
|Measr|+MAKALE                          |-MADDE      |-PUANLAYICI|Scale|
+-------------------------------------------------------------------------+
```

Şekil 1. Üç yüzey için Logit Cetvel

Şekil 1'in ilk sütununda logit-cetvelin ölçüm birimi (logit) gözlenmektedir. Çalışmada yer alan tüm düzeyler, bu ölçüm birimi üzerinden yorumlanır. Cetvelin ikinci sütununda ölçekler, belirlenmiş olan

ölçütleri karşılayabilme düzeyleri; yukarıdan aşağıya doğru en yüksekten en düşüğe doğru sıralanmıştır. Böylece, çalışmada esas alınan ölçütleri karşılayabilme düzeyi en yüksek 14 numaralı makalenin (1.13) ve en düşük 27 (-1.01) numaralı makalenin olduğu gözlenmektedir. Makalelerin düzeylerinin logit cetvelinde geniş bir aralıkta dağılım göstermiş olması birbirlerinden güvenilir bir şekilde ayrılabildiklerinin de bir göstergesi olmaktadır. Cetvelin üçüncü sütununda ise değerlendirme formunda yer alan ölçütlerin güçlük düzeylerine göre yukarıdan aşağıya doğru, güçlük düzeyi en yüksek olandan en düşük olana doğru sıralandığı görülmektedir. Buna göre, 16. ölçütün (*Ölçüt geçerliği analizi sonucu elde edilen değerler uygundur.*) güçlük düzeyi en yüksek (karşılanması en zor) (0.72) iken 3. ölçütün (Ölçeğin amacına uygun bir hedef kitle belirlenmiştir.) güçlük düzeyi en düşüktür (karşılanması en kolay) (-1.58). Şekil 1'in dördüncü sütununda puanlayıcılara ait ölçümler yer almaktadır. Bu sütunda yer alan veriler, en yüksek pozitif değere karşılık en katı puanlayıcı, en düşük negatif değere karşılık en cömert puanlayıcı gelecek şekilde yorumlanır. Böylece, çalışmada yer alan 4. puanlayıcının (0.35 logit) en katı, 1. puanlayıcının (-0.51 logit) ise en cömert puanlayıcı olduğu gözlenmektedir.

Logit cetvel, tüm yüzeylerin göreli durumlarının doğrusal bir metrik üzerinden incelenmesine imkan tanımakla birlikte, yüzeylerin her birine ilişkin daha detaylı bilgiler ölçüm raporları ile aşağıda yer alan tablolar üzerinden açıklanmıştır. Tablo 1'de makaleler için elde edilen ölçüm raporları sunulmuştur.

Tablo 2. Makalelere İlişkin Ölçüm Raporu

	Logit Ölçüsü	Std. Hata	Uyum-içi	Uyum-dışı	
Ortalama	0.00	0.14	1.02	0.91	
Standart Sapma	0.40	0.02	0.33	0.39	
Güvenirlik = 0.88	Ayırma Oranı = 2.65		Ki-kare = 444.9	sd = 56	p =0.00

Tablo 2'deki değerler incelendiğinde, makalelerin belirlenen ölçütlere sahip olma düzeyleri ortalamasının 0 ve standart sapmasının 0.40 logit puan olduğu gözlenmektedir. Uyum-içi ve uyum-dışı istatistikleri ise sırasıyla, 1.02 ve 0.91 olarak elde edilmiştir. Uyum istatistikleri, model ile veri arasındaki uyumun derecesini gösterir. Model ile veri arasında mükemmel bir uyum olduğunda bu istatistik değerleri 1 olmaktadır. Bunun yanı sıra; uyum-dışı istatistiği, beklenmeyen uç değerlere karşı uyum-içi istatistiğine göre çok daha duyarlıdır. Uyum-içi istatistiğinin 1'in üzerinde olması beklenenden daha yüksek, 1'den düşük olması ise beklenenden daha düşük bir varyansın gözlendiğine işaret eder (Güler, 2014; Hetherman, 2004). Alanyazında uyum istatistiklerine ilişkin kabul edilebilir değerlerin, 0.6 ile 1.5 (Lunz, Wright ve Linacre, 1990) ya da 0.5 ile 1.5 (Wright ve Linacre, 1994; Turner, 2003) arasında olabileceği belirtilmiştir. Buna göre, makale ölçüm raporunda yer alan uyum istatistikleri ortalamasının 1'e çok yakın olması, model-veri uyumunun oldukça yüksek olduğuna işaret etmektedir. Tablo 2'deki ayırma oranı ve güvenirlik değerleri incelendiğinde; bu değerlerin sırasıyla 2.65 ve 0.88 olduğu gözlenmektedir. Buradaki güvenirlik değeri, Cronbach alfa güvenirlik değeri gibi 0 ile 1 arasında bir değer alır ve benzer şekilde yorumlanır. Elde edilen değerin oldukça yüksek çıkması, makalelerin belirlenen ölçütler açısından yeteri derecede ayırt edilebildiğini göstermektedir.

Makalelerin değerlendirilmesinde belirlenen ölçütlere ilişkin ölçüm raporu sonuçları Tablo 3'te verilmiştir.

Tablo 3. Değerlendirme Ölçütlerine İlişkin Ölçüm Raporu

Ölçütler	Ölçüm	Std. Hata	Uyum İçi	Uyum Dışı	
Ö16	0.72	0.07	1.39	1.17	
Ö14	0.46	0.06	1.16	1.24	
Ö5	0.37	0.06	0.85	0.96	
Ö7	0.25	0.06	1.28	1.31	
Ö9	-0.07	0.06	1.10	0.73	
Ö15	-0.34	0.06	1.19	1.18	
Ö8	-0.74	0.07	0.93	0.87	
Ö11	-0.76	0.07	1.06	1.20	
Ö6	-0.87	0.08	0.73	0.75	
Ö4	-0.99	0.08	0.61	0.87	
Ö2	-1.12	0.09	0.55	0.80	
Ö17	-1.24	0.10	0.79	0.80	
Ö13	-1.35	0.11	0.95	1.07	
Ö12	-1.47	0.12	1.04	1.11	
Ö10	-1.53	0.12	1.49	1.16	
Ö1	-1.55	0.12	0.50	0.92	
Ö3	-1.58	0.13	0.68	0.93	
Ortalama	-0.69	0.09	0.96	0.91	
Standart Sapma	0.78	0.03	0.29	0.25	
Güvenirlik = 0.99	Ayırma Oranı = 8.69		Ki-kare = 1465.6	sd = 16	p = 0.00

Tablo 3'te yer alan ölçütlere ilişkin ölçüm raporu incelendiğinde, ölçütlerin güçlük düzeylerine göre en zor 0.72 logit değer ile Ö16 (*Ölçüt geçerliği analizi sonucu elde edilen değerler uygundur.*) ve en kolay -3.14 logit değer ile Ö3 (*Ölçeğin amacına uygun bir hedef kitle belirlenmiştir*.) ölçütünün olduğu gözlenmektedir. Bu ölçütlere ilişkin uyum-içi ve uyum-dışı istatistikleri sırasıyla 0.50 ile 1.49 ve 0.80 ile 1.24 arasında değişen değerler almaktadır. Uyum istatistikleri için istenilen değer aralığının 0.5 ile 1.5 olup bu aralık dışında kalan istatistikler zayıf uyuma işaret etmektedir (Turner, 2003; Wright ve Linacre, 1994). Buna göre, model ile veri uyumunu olumsuz etkileyen bir ölçütün bulunmadığı söylenebilir. Ölçütlere ilişkin güvenirlik değerinin 0.99 ile oldukça yüksek bir değer elde edilmesiyle de değerlendirmede kullanılan ölçütler, makalelerin belirlenen ölçütlere sahip olma düzeylerine göre ayırmada güvenilir bir şekilde işlev göstermişlerdir şeklinde yorumlanabilir.

Makaleleri belirlenen ölçütlere göre değerlendiren puanlayıcılara ilişkin ölçüm raporu sonuçları Tablo 4'te verilmiştir.

Tablo 4. Puanlayıcı Ölçüm Raporu

Puanlayıcılar	Ölçüm	Std. Hata	Uyum İçi	Uyum Dışı	
P4	0.35	0.03	1.00	1.06	
P3	0.15	0.04	0.86	0.75	
P2	0.01	0.04	1.07	0.94	
P1	-0.51	0.04	1.19	0.88	
Ortalama	0.00	0.04	1.03	0.91	
Standart Sapma	0.37	0.00	0.14	0.13	
Güvenirlik = 0.99	Ayırma indeksi = 9.77		Ki-kare = 260.4	sd = 3	p = 0.00

Tablo 4'te, puanlayıcıların ölçüm raporuna göre güvenirlik ve ayırma indeksi değerlerinin sırasıyla; 0.99 ve 9.77 olduğu gözlenmektedir. Oldukça yüksek olan güvenirlik değeri, puanlayıcıların puanlamalarının benzerliğinin değil farklılığının güvenirlik değeridir (İlhan, 2016). Buna ek olarak, yüksek olan ayırma indeksi de puanlayıcıların puanlamaları arasındaki farklılığın düzeyini ifade

eder. Puanlayıcılar tam bir tutarlılık ile puanlama yaptıklarında ayırma indeksinin alacağı değer 0 olmaktadır. Böylece, bu iki değer göz önüne alındığında, çalışmada yer alan puanlayıcıların puanlamadaki katılık ve cömertlik düzeylerinin farklılık gösterdiği yorumu yapılabilir. Tablo 4'teki ölçüm değerleri incelendiğinde, P4 numaralı puanlayıcının en katı ve P1 numaralı puanlayıcının ise en cömert puanlama yaptıkları gözlenmektedir. Ancak, gözlenen uyum-içi ve uyum-dışı istatistiklerinin kabul edilebilir aralık olan 0.5 - 1.5 değerleri arasında olması, puanlayıcıların puanlarının modelle uyumluğu olduğuna işaret eder (Wright ve Linacre, 1994). Buna göre, model-veri uyumunu olumsuz etkileyen bir puanlayıcının olmadığı söylenebilir.

Makale değerlendirme aşamasında kullanılan değerlendirme formunda yer alan ölçütlerin ölçüm raporu sonuçları Tablo 5'te verilmiştir.

Tablo 5. Değerlendirme Formu Ölçütlerinin Ölçüm Raporu

Puanlama düzeyleri	Frekans	%	Yığılmalı %	Ortalama Ölçüm	Beklenen Ölçüm	Uyum Dışı
0	823	21	21	-0.39	-0.37	0.70
1	193	5	26	0.11	0.09	0.70
2	445	11	38	0.74	0.62	1.10
3	2413	62	100	1.10	1.12	1.20

Tablo 5'te, değerlendirme formunda bulunan ölçütlerin frekans ve yüzde dağılımları incelendiğinde 0 - 3 arasında yapılan değerlendirmede "1" düzeyinin (bilgi var ama hiç açıklanmamış) en az kullanıldığı, bunu sırasıyla "2" (bilgi var ve kısmen açıklanmış) ve "0" (bilgi hiç yok) düzeylerinin takip ettiği gözlenmektedir. Değerlendirme formunda yer alan "3" düzeyinin (bilgi var ve gerekli şekilde açıklanmış) ise en çok kullanılan düzey olduğu belirlenmiştir. Düzeylerin işlevini yerine getirdiğinin söylenebilmesi için puanlama düzeylerindeki dağılımın dengeli olması ve her düzeyde en az 10 katılımcının bulunmuş olması ve uyum-dışı istatistiklerinin de 0.5 ile 1.5 aralığında değerler almış olması beklenir (İlhan, 2016; Linacre, 2014). Buna göre, düzeylere ilişkin frekans değerleri incelendiğinde tüm düzeyler için istenilen değerlere ulaşılabildiği ve uyum-dışı istatistiklerinin de istenilen aralıkta yer aldığı gözlenmektedir. Elde edilen bu sonuçlar, değerlendirme formunda yer alan ölçütlerin karşılanması beklenen varsayımları sağladığını, bir sorun olmadığını işaret etmektedir.

Makalelerin Belirlenen Ölçütleri Karşılama Düzeylerinin İncelenmesi

Verilerin ÇYRM ile analizi sonucunda makale yüzeyine ilişkin elde edilen logit ölçüm değerleri yukarıda sunulmuştur. Ancak Linacre (2014)'nin de ifade ettiği üzere, logit değerler ile rapor edilen makalelerin belirlenen ölçütleri karşılama düzeyleri tüm araştırmacılar ve okuyucular için anlaşılır olmayabilir. Bu nedenle, logit birim ile ifade edilen makale düzeyleri, değerlendirme formunda yer alan ölçüm düzeylerine dönüştürülmelidir. Tablo 6'da, alanyazında yer alan bu dönüştürme işleminin basamakları ayrıntılı olarak tanımlanmaya çalışılmıştır (İlhan, 2016; Linacre, 2014).

Tablo 6. Logit Birim ile İfade Edilen Makalelerin Düzeylerinin Değerlendirme Formundaki Düzeylere Dönüştürülmesi Süreci

Dönüşüm adımları	İşlemler
1. adım: En düşük logit değer ile en yüksek logit değer arasındaki fark hesaplanır (Bknz Şekil 1).	1.13 - (-1.01) = 2.14
2.adım: Değerlendirme formundaki değerlerin ranj değeri belirlenir.	3 – 0 = 3
3. adım: 2. adımda bulunan bu ranj değeri 1. adımda bulunan değere bölünür.	3 / 2.14 = 1.402
4. adım: 3. adımda bulunan değer, makalelere ilişkin elde edilen en düşük logit ölçümü ile çarpılır.	-1.01 * 1.402 = -1.416
5. adım: 4. adımda ulaşılan değeri değerlendirme formunun en küçük düzeyine (sıfıra) eşitleyecek sabit belirlenir.	Bu değer 1.416'dır.
6. adım: Her bir makale için elde edilen logit ölçümü 3.adımda elde edilen değer ile çarpılır ve bu değer ile 5. adımdaki değer toplanır.	(Logit puan *1.402) + 1.416

Tablo 6'da yer alan süreç sonrası, incelenen makalelerden sadece birinin (% 1.7) değerlendirilme düzeyinin 3 olduğu gözlenmiştir. Bunu sırasıyla, 2.58 ile 2.20 arasında değerlendirilen 5 makale (% 8.8), 1.91 ile 1.01 arasında değerlendirilen 41 makale (% 72) izlemektedir. En düşük değerlendirme düzeyi olan 0.95 ile 0 arasında ise 10 makalenin (% 17.5) yer aldığı belirlenmiştir. Tablo 6'da yer alan adımlar, makale değerlendirme formunda yer alan değerlendirme ölçütleri için tekrarlanmış ve ölçüt düzeylerinin KTK'daki madde güçlük düzeyleri gibi 0 ile 1 arasındaki değerleri belirlenmiştir. Ancak burada KTK'dan farklı olarak 1'e en yakın ölçüt en zor, 0'a en yakın ölçüt ise en kolay ölçüt olarak yorumlanmaktadır. Ölçütlere ilişkin elde edilen bu değerler Tablo 7'de verilmiştir.

Tablo 7. Ölçütlerin Logit Değerlerinin [0, 1] Aralığındaki Değerleri

No	Ölçüt ifadeleri	Dönüştürülmüş değerler
1	Ölçülmesi amaçlanan değişken açıkça belirtilmiştir.	0.02
2	Yeni bir ölçeğe gereksinim duyulmasının nedenleri gerekçeleriyle belirtilmiştir.	0.21
3	Ölçeğin amacına uygun bir hedef kitle belirlenmiştir.	0.01
4	Madde havuzu uygun yöntemler kullanılarak oluşturulmuştur (Literatür tarama, Uzman görüşü alma, Kompozisyon yazdırma vb...)	0.26
5	Deneme uygulaması için üretilen madde sayısı, nihai ölçekte bulunması öngörülen madde sayısından daha fazla (en az 2-3 katı) olacak şekildedir.	0.85
6	Madde havuzu oluşturulduktan sonra, alan uzmanları ve ölçme ve değerlendirme uzmanlarından oluşan bir grup uzman tarafından maddeler gözden geçirilmiştir.	0.32
7	Oluşturulan madde havuzuna ön deneme uygulaması yapılmıştır.	0.80
8	Deneme uygulamasının örneklemi, ölçeğin amacı doğrultusunda ölçülecek özelliğe sahip olan bireylerin tümünü temsil edici niteliktedir.	0.37
9	Örneklem büyüklüğünün belirlenmesinde ölçek ve madde analizinde yapılacak olan analizler dikkate alınmıştır.	0.66
10	Elde edilen veriler faktör analizine uygundur.	0.03
11	Belirlenen faktörlerin öz değerleri 1'den büyüktür.	0.36
12	Ölçeğin açıklanan varyansı %40'tan yüksektir (tek faktörlü ölçek için %30'dan büyük).	0.06
13	Seçilen maddelerin faktör yük değerleri 0.30'dan büyüktür.	0.11
14	Doğrulayıcı faktör analizi için toplanan veriler ile ilgili betimsel bilgiler verilmiştir.	0.89
15	DFA analizi sonucunda elde edilen model veri uyumu indeksleri uygun değerlerdir.	0.54
16	Ölçüt geçerliği analizi sonucunda elde edilen değerler uygundur.	1.00
17	Ölçeğin güvenirlik analizi sonucunda elde edilen değerler ölçeğin tamamı ve tüm alt boyutları için uygundur.	0.16

Tablo 7'de yer alan ölçütlerin [0, 1] aralığındaki değerleri incelendiğinde, değerlendiriciler tarafından makalelerin en büyük çoğunluğu tarafından istenilen düzeyde olduğu belirtilen ölçütün 3. ölçüt (Ölçeğin amacına uygun bir hedef kitle belirlenmiştir.) olduğu, bunu sırasıyla 1. (Ölçülmesi amaçlanan değişken açıkça belirtilmiştir.) ve 10. (Elde edilen veriler faktör analizine uygundur.) ölçütlerinin izlediği gözlenmiştir. Bunun yanı sıra, değerlendiriciler tarafından makalelerin büyük bir çoğunluğu tarafından istenilen düzeyin karşılanmadığı belirtilen ölçütün ise 16. ölçüt (Ölçüt geçerliği analizi sonucunda elde edilen değerler uygundur.) olduğu, bunu sırasıyla 14. (doğrulayıcı faktör analizi için toplanan veriler ile ilgili betimsel bilgiler verilmiştir.) ve 5. (Deneme uygulaması için üretilen madde sayısı, nihai ölçekte bulunması öngörülen madde sayısından daha fazla (en az 2-3 katı) olacak şekildedir.) ölçütlerin izlediği belirlenmiştir.

Çalışmada son olarak, incelenen 57 makalede, geliştirilen ölçeklere yer verilme durumu ile puanlamaya ilişkin bilginin verilme durumları gözden geçirilmiştir. Bu bağlamda makalelerin 27 tanesinde (%47,4) geliştirilen ölçekler makale içeriğinde verilirken 30 makalede (%52,6) bu durum gözlenmemiştir. Geliştirilen ölçeklerin diğer araştırmacılar tarafından kullanıldığında puanlamanın nasıl yapılacağına dair bilginin verilme durumu ise makalelerin %56,1'inde (32 makale) gözlenirken % 43,9'unda (25 makale) puanlama ile ilgili bir bilgiye rastlanmamıştır.

SONUÇLAR ve TARTIŞMA

Bu araştırmada, Türkiye'deki eğitim alanında SSCI dizininde yer alan üç dergide 2010-2015 yılları arasında yayınlanan 57 ölçek geliştirme makalesi, ÇYRM ile analiz edilerek incelenmiştir. Çalışma sonucu ulaşılan bulgularda, makalelerin güvenilir bir şekilde ayırt edilebildiği, belirlenen ölçütlerin güvenilir olduğu ve ölçüt kategorilerinin uygun ve yeterli ölçüm yapabildiği gözlenmiştir. Çalışmada yer alan dört puanlayıcının, makalelere verdikleri puanların, katılık-cömertlik düzeylerinin birbirinden farklı olduğu ancak model-veri uyumunun sağlandığı belirlenmiştir. Alan yazın incelendiğinde, farklı puanlayıcıların yer aldığı çalışmalarda, puanlayıcıların puanlama öncesi bilgilendirilme sürecinden, farklı eğitim süreçlerinden geçmiş olmaları ve farklı eğitim düzeylerinde olmalarından kaynaklanan puanlayıcılar arası farklılıkların olabileceği vurgulanmaktadır (Baştürk ve Işıkoğlu, 2008; Mulqueen, Baker ve Dismuskes, 2000; Semerci, 2012; Semerci, 2011). Ayrıca, ÇYRM analizinde puanlayıcılar arası farklılığın olması bir sınırlılık olmamakta; bu farklılıklar kontrol altında tutularak birey ve maddelere (bu çalışma için makaleler ve ölçütlere) ilişkin istatistikler bu şekilde hesaplanmaktadır (Abu Kassim, 2007).

Analizler sonucu elde edilen bulgulara göre, çalışmada kullanılan formdaki 17 ölçütün tümü dikkate alındığında 57 makaleden sadece bir tanesinde, belirlenen bütün ölçütlere ilişkin bilginin bulunduğu ve gerekli açıklamaları kapsadığı belirlenmiştir. İncelenen makalelerden beşinde (% 8.8), ölçütlere ilişkin bilginin bulunduğu ve kısmen açıklama yapıldığı, 41'inde (% 72) sadece ölçütlere ilişkin bilginin yer aldığı ancak her hangi bir açıklama yapılmadığı ve 10'unda (% 17.5) ise ölçütlerle ilgili hiç bir bilginin yer almadığı belirlenmiştir. Alan yazın incelendiğinde, Türkiye'deki ölçek geliştirme makalelerinin incelendiği diğer çalışmalarda da benzer sonuçlara rastlanmaktadır. Acar Güvendir ve Özer Özkan (2015,) çalışmalarında toplam 26 ölçek geliştirme makalesi incelemişlerdir ve çalışmalarında yer alan "ölçek geliştirmede izlenmesi gereken 22 aşamadan" sadece sekizinin (%36.4), bu makalelerin %75'i ve daha fazlası tarafından sağlanabildiği belirlenmiştir.

Çalışmada kullanılan makale değerlendirme ölçütleri detaylı incelendiğinde, üç ölçütün incelenen makalelerin çoğunda oldukça yüksek düzeyde olduğu sonucuna ulaşılmıştır. Çalışmalarda kullanılan ölçeklerin çok farklı hedef kitleleri olmakta ve bu sebeple aynı değişken için dahi farklı hedef kitleleri için ölçeklerde yer alan ifadelerde farklılaşmaktadır. Bu sebeple ölçek geliştirme çalışması yürüten bir araştırmacının önceliklerinden birinin amacına uygun hedef kitlesini belirlemek olması gerekmektedir. Bu gereklilikten dolayı da incelenen çalışmalarda bu ölçütün karşılanma düzeyi oldukça yüksek çıkmıştır. Bunun yanı sıra, ölçek geliştirme çalışması yürüten araştırmacıların, araştırmalarına başlamadan önce, ölçme aracı geliştireceği değişken ile ilgili yeterli bilgiye sahip olması gerekmektedir. Bu sebeple yürüttüğü çalışmada da bu değişken ile ilgili bilgi vermesi kaçınılmazdır. Bu durumun yansıması, incelenen makalelerde ölçülmesi amaçlanan değişkenin

açıklanma oranının oldukça yüksek çıkması bulgusu gözlenmektedir. Son olarak, incelenen 57 makalenin verilerinin faktör analizine uygunluğunun sağlanması ölçütü oldukça yüksek bir oranda gözlenmiştir. Ölçek geliştirme çalışmalarında araştırmacıların neredeyse tamamının topladıkları verilere faktör analizi yaptıkları gözlenmektedir. Toplanan veriler, faktör analizi yapmaya uygun değilse, çalışmanın ilerlemesi ve dolayısıyla bir yayına dönüşmesi de mümkün olmayacaktır. Bu ölçütün karşılanma oranının yüksek olmasının gerekçesi bu şekilde açıklanabilir.

Çalışmada kullanılan 17 makale değerlendirme ölçütünden yukarıda belirtildiği üzere oldukça yüksek oranlarda karşılanan ölçütlerin yanı sıra değerlendiriciler tarafından istenilen düzeyde karşılanamayan ölçütlere de rastlanmıştır. Bunların ilki ölçüt geçerliği ile ilgili olan ölçüttür. Bu değerlendirme ölçütünün karşılanma oranının düşük olmasının sebebi, araştırmacıların ölçek geliştirme esnasında böyle bir adımdan haberdar olmamalarından çok başka nedenlere dayandırılabilir. Öncelikle ölçek geliştirme çalışması yürüten araştırmacının, ölçüt geçerliği ile ilgili veri toplayabilmesi için öncelikle ölçek geliştirdiği değişken ile ilişkili bir başka ölçek olması gerekmektedir. Böyle bir ölçeğin bulunması her zaman olası olmayabilir. Bunun yanı sıra ölçüt bulunsa dahi ölçüt geçerliğinde kullanılacak ölçeğin kullanılarak yeniden veri toplanması zaman, maliyet vb. sebeplerden ötürü tercih edilmemiş olabilir. İstenilen düzeyde karşılanamayan bir diğer ölçüt ise DFA analizi için toplanan veriler ile ilgili betimsel bilginin verilme durumudur. İncelenen çalışmalarda da görüldüğü üzere, araştırmacıların büyük bir çoğunluğu topladıkları veriler üzerinde sadece açımlayıcı faktör analizi (AFA) yapmakta ancak doğrulayıcı faktör analizi yürütmemektedirler. Bu durumun ilk gerekçesi, araştırmacının böyle bir gerekliliğin farkında olmamasından kaynaklanabilir. Çünkü alanyazındaki pek çok çalışmada sadece AFA yapılmakta ve araştırmacılarda bu çalışmaları örnek alarak çalışmalarını yürütmüş olabilirler. Bir diğer gerekçe ise AFA ve DFA'nın farklı örneklemlere uygulanması gerekliliğidir. Geliştirilmeye çalışılan ölçeğin uygulanacağı uygun hedef kitlesinden bireylere ulaşılması bir diğer ifadeyle yeterli sayıda kişiden veri toplanması oldukça zahmetli bir süreçtir. AFA'nın yanı sıra DFA'da yaparak bu süreci daha fazla zorlaştırmamak adına araştırmacılar, bu adımı atlamış olabilirler. Bir diğer gerekçe ise AFA'ya kıyasla DFA'nın araştırmacılar tarafından çokta sık kullanılmayan bilgisayar programlarında yapılması olabilir. Araştırmacılar kendilerine yabancı ve bu sebeple daha zor gelen bu adımı atlıyor olabilirler. Son olarak, incelenen makalelerde karşılanma oranı düşük olan ölçüt, deneme uygulamasında üretilen madde sayısının nihai ölçektekinin 2-3 katı olması gerekliliğidir. Normal şartlar altında herhangi bir davranışı yoklamaya yönelik madde yazmak zahmetli ve detaylı bir iştir. Aynı davranışını ölçen birden fazla madde yazma süreci ise daha da zordur. Bu sebeple araştırmacılar her zaman bir davranışla ilgili birden fazla madde yazmakta zorlanabilirler. Bu durumda da deneme formundaki madde sayısının nihai ölçekte bulunması amaçlanan madde sayınının 2-3 katı olması gerekliliği karşılanamayabilir. Bu durumun bir diğer gerekçesi ise deneme formundaki madde sayısındaki artış, ölçeğin cevaplanma süresini ve dolayısıyla cevaplayıcıların konsantrasyonlarını etkileyeceği için araştırmacılar kasıtlı olarak daha az sayıda maddeden oluşan deneme formları geliştirmiş olabilir. Çalışma kapsamında son olarak makalelerde bulunma durumu incelenen iki madde yer almaktadır. Bunlardan ilki geliştirilen ölçeğin çalışmalarda verilme durumudur. Elde edilen sonuçlarda, geliştirilen ölçeğin verilme durumu yaklaşık %50 civarında gözlenmiştir. Geriye kalan makalelerde verilmeme nedeni, makalelerdeki sayfa sınırından ölçeklerin makale kapsamında yer alamaması olabileceği gibi ölçeği geliştiren çalışmacıların, geliştirdikleri ölçeği kullanan bireylerden anında haberdar olma istekleri ve kendilerinden izin alındıktan sonra ölçeği diğer araştırmacılara gönderme istekleri olabilir. Son olarak incelenen ikinci madde ise geliştirilen ölçeklerle ilgili puanlama bilgisinin verilme durumudur. İncelenen makalelerin yaklaşık yarısında (%56,1) puanlama ile ilgili bilgi bulunmaktadır. Puanlama bilgisi verilmeyen makalelerdeki bu bilgi eksikliği, geliştirilen ölçekleri çalışmalarında kullanacak diğer araştırmacıların elde ettikleri puanların yorumlamalarını güçleştirecektir. Bu sebeple, çalışmalarda ölçeklerden alınacak en yüksek ve en düşük puanların yanı sıra gerekiyorsa belirli aralıkların tanımlanması yoluyla ölçeğin hem alt boyutları hem de bütünü itibariyle alınacak puanların yorumu ile ilgili bilgi verilmesi gerekmektedir.

KAYNAKÇA

Abu Kassim, N. L. (2007, June). Exploring rater judging behaviour using the many-facet Rasch model. *Paper Presented in the Second Biennial International Conference on Teaching and Learning of English in Asia: Exploring New Frontiers (TELiA2)*, Holiday Villa Beach & Spa Resort, Langkawi. Faculty of Communication and Modern Languages, Universiti Utara Malaysia.

Acar Güvendir, M. ve Özer Özkan, Y. (2015). Türkiye'deki eğitim alanında yayımlanan bilimsel dergilerde ölçek geliştirme ve uyarlama konulu makalelerin incelenmesi. *Elektronik Sosyal Bilimler Dergisi, 14*(52), 23-33.

Anshel, M. H., Weatherby, N. L., Kang M. ve Watson, T. (2009). Rasch Calibration Of A Unidimensional Perfectionism Inventory For Sport. *Psychology of Sport and Exercise, 10*(2009), 210-216.

Barrett, G. V. (1972). Research models of the future for industrial and organizational psychology. *Personnel Psychology, 25, 1-17*

Baştürk, R. ve Işıkoğlu, N. (2008). Okul Öncesi Eğitim Kurumlarının İşlevsel Kalitelerinin Çok-Yüzeyli Rasch Ölçme Modeli İle Analizi. *Kuram ve Uygulamada Eğitim Bilimleri, 8*(1), 7-32.

Boztunç Öztürk, N., Eroğlu, G. ve Kelecioğlu, H. (2014). Eğitim alanında yapılan ölçek uyarlama makalelerinin incelenmesi. *Eğitim ve Bilim, 40*(178), 123-137.

Brinthaupt, T. M. ve Kang, M. (2012). Many-Faceted Rasch Calibraton: An Example Using the Self-Talk Scale. *Assessmen*t, *21*(2), 241-249.

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment, 7*, 309-319.

Cohen R. J. & Swerdlik M.E. (2010). *Psychological testing and assessment.* Boston: McGraw-Hill Companies.

Cook, J. D., Hepworth, S. J., Wall, T. D., & Warr, P. B. (1981). *The experience of work*. London: Academic Press.

Çüm, S. ve Koç, N. (2013). Türkiye'de psikoloji ve eğitim bilimleri dergilerinde yayımlanan ölçek geliştirme ve uyarlama çalışmalarının incelenmesi. *Eğitim Bilimleri ve Uygulama, 12*(24), 115-135.

Delice, A. ve Ergene, Ö. (2015). Investigation of scale development and adaptation studies: An example of mathematics education articles. *Karaelmas Journal of Educational Sciences 3, 60-75*

DeVellis, R.F. (2003). *Scale development: Theory and applications.* Newbury Park: Sage Publications, Inc.

Edenborough R. (1999). *Using psychometrics: a practical guide to testing and assessment.* London: Kogan Page.

Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement, 31*(2), 93-112.

Erkuş, A. (2007). Ölçek geliştirme ve uyarlama çalışmalarında karşılaşılan sorunlar. *Türk Psikoloji Bülteni, 13*(40), 17-25.

Erkuş, A. (2012). *Psikolojide ölçme ve ölçek geliştirme-1: Temel kavramlar ve işlemler.* Ankara: Pegem Akademi.

Güler, N. (2014). Analysis Of Open-Ended Statistics Questions With Many Facet Rasch Model. *Eurasian Journal of Educational Research, 55*, 73-90.

Güler, N. ve Gelbal, S. (2010). A Study Based On Classical Test Theory And Many Facet Rasch Model. *Eurasian Journal of Educational Research, 38*, 108-125.

Hetherman, S. C. (2004). *An application of multi faceted Rasch measurement to monitör effectiveness of the written composition in English in the new york city department of education.* Unpublished phd. dissertation. Teacher College, Colombia University, Colombia.

Hinkin, T. R. (1995). A review of scale development practices in the study of organizations. *Journal of Management, 21*(5), 967-988.

Hinkin, T. R., Tracey, J. B. & Enz, C.A. (1997). Scale construction: developing reliable and valid measurement instruments. *Journal of Hospitality and Tourism Research, 21*(1), 100-120.

İlhan, M. (2016). Açık uçlu sorularla yapılan ölçmelerde klasik test kuramı ve çok yüzeyli Rasch modeline göre hesaplanan yetenek kestirimlerinin karşılaştırılması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 31*(2), 346-368. Doi: 10.16986/HUJE.2016015182

Karasar, N. (1998). *Araştırmalarda rapor hazırlama yöntemi.* Ankara: Pars Matbaacılık Sanayi.

Kim, Y., Park, I. & Kang, M. (2012). Examining rater effects of the TGMD-2 on children with intellectual disability. *Adapted Physical Activity Quarterly, 29*, 346-365.

Lee, M. Peterson, J. J. ve Dixon, A. (2010). Rasch Calibration Of Physical Activity Self-Efficacy And Social Support Scale For Persons With İntellectual Disabilities. *Research in Developmental Disabilities, 31*(2010), 903-913.

Linacre, J.M. (2014). *A user's guide to FACETS Rasch-model computer programs.* [Available online at: http://www.winsteps.com/a/facets-manual.pdf], Retrieved on July 13, 2015.

Linacre, J. M. (2007). *A User's Guide to FACETS: Rasch Model Computer Programs.* Chicago, IL.

Linacre, J. M. (1989). *Many-facet Rasch Measurement.* Yayınlanmamış doktora tezi. University of Chicago, Chicago, IL.

Lunz, M. E., Wright, B. D. & Linacre, J. M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education, 3*(4), 331-345.

Mulqueen C., Baker D., & Dismukes, R. K. (2000, April). Using multifacet Rasch analysis to examine the effectiveness of rater training. *Presented at the 15th Annual Conference for the Society for Industrial and Organizational Psychology (SIOP). New Orleans*

Murphy K.R. & Davidshofer C.O. (2005). *Psychological testing: Principles and applications.* New Jersey: Pearson Education International.

Randall, J. & Engelhard, G. Jr. (2009). Examining teacher grades using Rasch measurement theory. *Journal of Educational Measurement, 46*(1), 1-18.

Revesz, A. (2012). Working Memory and the Observed Effectiveness of Recasts on Different L2 Outcome Measures. *Language Learning, 62*(1), 93-132.

Schriesheim, C. A., Powers, K. J., Scandura, T. A., Gardiner, C. C., Lankau, M. J. (1993). Improving construct measurement in management research: Comments and a quantitative approach for assessing the theoretical content adequacy of paper-and-pencil survey-type instruments. *Journal of Management, 19, 385-417.*

Semerci, Ç. (2011). Doktora yeterlikler çerçevesinde öğretim üyesi, akran ve öz değerlendirmelerin rasch ölçme modeliyle analizi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 2(2), 164-171.*

Semerci, Ç. (2012). Öğrencilerin BÖTE Bölümüne İlişkin Görüşlerinin Rasch Ölçme Modeline Göre Değerlendirilmesi (Fırat Üniversitesi Örneği). *E-Journal of World Science Academy, NWSA-Education Sciences, 1CO542, 7(2), 777-784.*

Smith, V. E. & Kulikowich, M. J. (2004). An application of generalizability theory and many facet Rasch measurement using a complex problem-solving skills assessment. *Educational and Psychological Measurement, 64*, 617-639.

Stone, E. (1978). *Research methods in organizational behavior*. Glenview, IL: Scott, Foresman.

Sudweeks, R.R., Reeve, S., & Bradshaw, W.S. (2005). A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing*, 9(3), 239-261.

Tavşancıl, E., Güler, G. ve Ayan, C. (2014). *2002-2012 yılları arasında Türkiye'de geliştirilen bazı tutum ölçeği geliştirme çalışmalarının ölçek geliştirme süreci açısından incelenmesi.* IV. Ulusal Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongresi (Uluslararası Katılımlı) 9-13 Haziran, Hacettepe Üniversitesi, Ankara.

Tezbaşaran, Ata. (2008). *Likert tipi ölçek geliştirme kılavuzu*. Üçüncü baskı, Türk Psikologlar Derneği Yayını, Ankara.

Turner, J. (2003). *Examining on art portfolio assessment using a many facet Rasch measurement model*. Unpublished phd. dissertation. Boston College, Boston.

Wright, B. D. ve Linacre, J. M. (1992). Combining and splitting categories. *Rasch Measurement Transactions, 6*(3) 233-235.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement: Transactions of the Rasch Measurement SIG*, 8(3), 370.

Yıldırım, A. ve Simsek, H. (2008). *Sosyal bilimlerde nitel araştırma yöntemleri*. Ankara: Seçkin Yayıncılık.

EXTENDED ABSTRACT

Introduction

The main reason for using scales as measurement instruments is the intention to measure the variables that are regarded that they exist in education and psychology but which cannot be directly observed. Such variables that are considered to exist in individuals are called structures and those structures cause visible behaviors in people. Scales aim to determine the extent to which individuals have the structures by using rubrics.

The generalizability, functionality and firmness of the findings obtained by applying the scales to individuals are related with the reliability and validity of those instruments. The results obtained from studies using the same scale or from studies using different scales developed to measure the same properties, for instance, can be inconsistent. It is thought that this situation may basically stem from the fact that measurement tools yielding reliable and valid results have not been used.

Therefore, the quality of those tools used to measure the directly unobservable variables is very important.

An increase has been observed in scale development and adaptation studies especially in recent years. One reason of the increase is the pressure about doing research that academicians feel over them. This, in turn, results in low quality studies. On examining the studies performed, it is observed that there are a number of problems and inadequacies in the studies performed or in the process of conducting the studies.

This study aims to demonstrate the extent to which the studies on scale development published in scientific journals appearing in SSCI in the period between 2010 and 2015 took the necessary steps in the "process of scale development". For this purpose, the data collected were primarily analyzed for each facet in terms of reliability by using many-faceted Rasch Model, situations appearing to be problematic were revised, and thus the data obtained were put to analysis. In this way, the levels and reliability of the facet of raters assessing the articles, the criteria used in rating and the articles considered were identified. Then, an attempt was made in this study to describe the strengths and weaknesses of the articles on "scale development" in terms of the criteria set. It is hoped that this study will contribute significantly to the literature by analyzing all the facets through many-faceted Rasch model and by determining the place of the articles considered and the criteria set in equal interval scale in addition to offering data supportive of similar studies evaluating scale development articles. This study also aims to set an example demonstrating that many-faceted Rasch model can be used in similar situations of measurement.

Method

This study analyses the articles concerning scale development and published in the journals of education surveyed in SSCI (Education and Science, Journal of Hacettepe University Faculty of Education, Journal of Educational Sciences in Theory and Practice) in the period between 2010 and 2015. Studies developing scales only for measuring typical behaviors and studies reporting the process of developing such scales have been included in this study. Thus, 57 articles in total whose full texts are accessible in three journals of education constitute the sample of this study.

This study reviews the references published in relation to the steps to be followed in and necessity of scale development as well as the studies analyzing scale development articles. A "Scale Development Article Investigation Form" has been developed by the researchers considering the general tendency in the literature. The form contains items, and the first 17 items are the criteria that can be scored between 0 and 3. Scoring in those criteria is described as "(0) no information; (1) information available but not explained; (2) information available and partly explained; (3) information available and explained in an appropriate way. The other two items are yes-no questions asking whether or not the scale developed was available in the studies and whether or not information on scoring the scale is available.

Results and Discussion

The findings demonstrated that the articles had been discriminated in a reliable way, the criteria set were reliable and the categories of the criteria were able to measure appropriately and adequately. It was found that the four raters available in the study differed in their levels of strictness and generosity in rating the articles but model-data fit had been attained. The findings also showed that only one out of 57 articles included information on the criteria set and offered the necessary explanation considering the 17 criteria in the form used in this study. Accordingly, five of all the studies (8.8%) contained information about the criteria set and offered partial explanations, 41 of them (72%) contained information on the criteria set but offered no explanations on them, and 10 of the studies (17.5%) contained no information on the criteria set. The 14th and 16th criteria which are related to descriptive information about the data collected for DFA and criterion validity respectively did not met adequately Finally, the criterion met at low levels in the articles considered was the necessity that the number of items created in the trial application should be two or two times more than in the final scale.