Hacettepe Üniversitesi Eğitim Fakültesi Dergisi (H. U. Journal of Education) Özel Sayı 2: 264-271 [2012]

ÖLÇEK GELİŞTİRME ÇALIŞMALARINDA KAPSAM GEÇERLİK ÖLÇÜLERİ: KAPSAM GEÇERLİK İNDEKSİ VE KAPPA İSTATİSTİĞİNİN KARŞILAŞTIRILMASI*

CONTENT VALIDITY MEASURES IN SCALE DEVELOPMENT STUDIES: COMPARISON OF CONTENT VALIDITY INDEX AND KAPPA STATICS

Halil YURDUGÜL**, Fatma BAYRAK***

ÖZET: Ölçme araçlarının geliştirilmesi iki aşamada ele alınabilir. Bunlar sırasıyla; ölçme aracının tasarımı ve pilot uygulama aşamasıdır. Her iki aşamaya özgü geçerlik kavramları söz konusudur. Özellikle ölçme aracının tasarımı aşamasında kapsam ve görünüş geçerliği ön plana çıkarken, pilot uygulama aşamasında ise yapı ve ölçüt geçerlikleri ele alınmaktadır. Özünde ise; kapsam geçerliği aynı zamanda yapı geçerliği için bir ön koşuldur. Kapsam geçerlikleri ölçmeye konu olan alandaki uzman görüşlerinden elde edilmektedir. Ancak nitel olarak ele alınan bu görüşlerin raporlama ve iletişim güçlüğünden dolayı nicel ifadelere dönüştürme çalışmaları vardır. Bunlardan öne çıkan ilk ikisi ise kapsam geçerlik (oranı) indeksi ve kappa istatistiğidir. Bu çalışmada her iki kapsam geçerlik ölçüsünün tutarlığı ele alınmıştır. Uygulamada bir ölçek geliştirme çalışmasının tasarım aşamasında 10 uzman görüşü alınmış ve kapsam geçerlik değerleri bu ölçülere göre elde edilmiştir. Daha sonra ise uygulama aşamasında faktör analizinden elde edilen faktör yük değerleri ve ortalama açıklanan varyans değerleri ile tutarlıkları araştırılmıştır. Elde edilen bulgular tartışılmıştır.

Anahtar sözcükler: Kapsam geçerliği, ölçek geliştirme, kapsam geçerlik indeksi, kappa istatistiği

ABSTRACT: It can be stated that there are two phases in the scale development process. The first phase is designing the scale; and the other phase is pre-application. There are different validity concepts specific to each phase. Particularly, while content and face validities have leading role in designing scale, the construct and criterion validities have come to the forefront in pre-application phase. In fact, content validity is prerequisite for construct validity. Content validity is often established through qualitative expert reviews. Content validity information often derives from reviews to be undertaken by subject matter experts as qualitative data. However, there are some studies about transforming these qualitative data into quantitative form; because it is difficult to report qualitative reviews of experts. Content Validity Index (CVI) and kappa statistics have come to the forefront in the studies. The purpose of the present study is to examine the consistency of Content Validity Index (CVI) and kappa statistics. For the purpose, in designing phase for a developing scale, data obtained from ten expert judges were analyzed with two approaches. In pre-application, the consistency of Content Validity Index (CVI) and kappa statistics with average variance extracted (AVE) and factor loadings obtained from factor analysis were examined.

Keywords: Content validity, construct validity, scale development, content validity index, kappa statistics

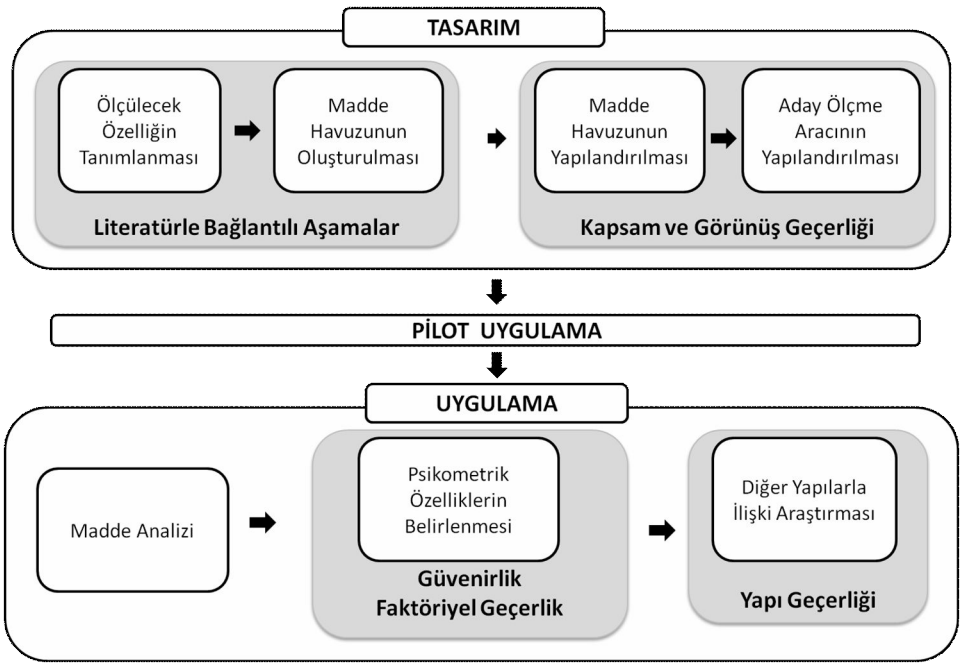
1. GİRİŞ

Eğitim ve psikoloji alanındaki araştırmalarda ölçmelerin güvenirliği ve geçerliği önemli konuların başında gelmektedir. Güvenirlik, ölçmelerin turtalılığı ile ilgili bir kavram iken geçerlik ise ölçmelerin amaçlanan yapıyı ölçme derecesi olarak ifade edilebilir. Ancak bu iki konu özellikle ölçmeleri elde etmekte kullanılan ölçme araçlarının geliştirilme sürecinde yer alır. Bilindiği gibi; ölçme araçlarının geliştirilmesi iki aşamada ele alınabilir. Bunlar sırasıyla; ölçme aracının tasarımı ve pilot uygulama aşamasıdır. Her iki aşamaya özgü birbirinden farklı geçerlik kavramları söz konusudur. Özellikle ölçme aracının tasarımı aşamasında kapsam ve görünüş geçerliği ön plana çıkarken, uygulama aşamasında ise faktöriyel, ölçüt (yordama, uygunluk) ve yapı (yakınsama ve ıraksama) geçerlikleri ele alınır. Şekil 1'de ölçek geliştirme süreci bir şema olarak verilmiştir. Bu sürecin ilk aşaması ölçme aracının tasarım aşaması olarak adlandırılabilir. Bu aşama, ölçme aracının ölçmesi amaçlanan eğitimsel-psikolojik yapının tanımlanması, bu yapıya ilişkin görgül gösterge kümesinin (madde havuzu) belirlenmesi ve aday ölçme aracının (yönerge ve seçenekleriyle birlikte bir) taslağının

* Bu çalışma 15-17 Mayıs tarihinde düzenlenmiş olan "The International Conference on Interdisciplinary Research in Education" isimli konferansta sunulmuştur.

** Doç. Dr., Hacettepe Üniversitesi, e-posta: yurdugul@hacettepe.edu.tr
*** Araş. Gör. Hacettepe Üniversitesi, e-posta: fbayrak@hacettepe.edu.tr

oluşturulma aşamasıdır. Bu aşamada ilgili literatüre ve konu alanı uzmanlarının görüşlerine sıkça başvurulur. Uzman görüşleri, özellikle ölçme aracı taslağının kapsam ve görünüş geçerliği ile ilgili bir sürece karşılık gelir. Kapsam geçerliği, ölçülecek olan yapının görgül göstergelerinin (ölçek maddeleri) ölçülmek istenilen yapıyı ne derece de temsil ettiğinin belirlenmesidir (Nunnally, 1978; Carmines & Zeller, 1979; Kerlinger, 1986; Pedhazur & Schmelkin, 1991). Anastasia (1988) kapsam geçerliğini görgül göstergelerin ilgili yapıyı ölçüp ölçmediği ile ilgili olduğunu, görünüş geçerliğinin ise görgül göstergelerin ilgili yapıyı ölçüyor görünmesiyle ilgilendiğini ifade etmiştir; bununla birlikte kapsam ve görünüş geçerliğinin kavram olarak birbirine benzerliklerinden dolayı kafa karıştırıcı olduğunu dile getirmiştir. Bazı araştırmacılar kapsam ve görünüş geçerliğinin farklı geçerlik türleri olduğunu vurgularken (DeVellis, 1991; Kerlinger, 1986) bazı araştırmacılar ise (Carmines and Zeller, 1979; Nunnally, 1978) görünüş geçerliğini kapsam geçerliğinin dolaylı bir değerlendirilmesi olduğunu belirtmişlerdir. Bu tartışmalara ek olarak, kapsam geçerliğinin (O'Leary-Kelly & Vokurka, 1998) ve görünüş geçerliğinin (Turner, 1979) yapı geçerliğinin bir ön koşulu olduğu ifade edilebilir. Aslında O'Leary-Kelly ve Vokurka (1998) a) önemli olan geçerlik türünün yapı geçerliği olduğunu, b) yapı geçerliğinin ise bir süreç olduğunu ve c) bu sürecin ilk adımının ise kapsam geçerliği olduğunu belirtmişlerdir. Bir bakıma; iyi yapılandırılmış ve işletilmiş bir kapsam geçerliği araştırmasının öncelikle ölçme aracının yapı geçerliğine hizmet edeceği açıktır.



Şekil 1: Ölçek geliştirme süreci ve aşamaları

Ölçek geliştirme sürecinin tasarım aşamasında ele alınan geçerlik türleri konu alanı uzmanlarının görüşleri üzerine dayanmaktadır. Ancak uzman görüşlerinin raporlanması ve iletişimindeki zorluklardan dolayı bu uzman görüşlerinin sayısal ifadelere dönüştürme çalışması ağırlık kazanmış, bu bağlamda bazı kapsam geçerlik ölçüleri önerilmiştir. Bunlardan bazıları Cohen (1960) tarafından geliştirilen kappa istatistiği, Tinsley–Weiss'in T indeksi (Tinsley & Weiss, 1975), James, Demaree ve Wolf (1993) tarafından önerilen rWG indeksi ile rWG(J) indeksleri ve Lindell, Brandt ve Whitney (1999) tarafından önerilen modifiye edilmiş rWG(J) indeksleridir. Bunlardan bir kısmı, ilgili maddeye kapsam geçerliği için "uygun" diyen uzmanların oranlarını ele alırken bir kısmı da uzman görüşlerinin uyumu üzerine kuruludur. Bu araştırma, kapsam geçerlik oranı/indeksi ölçüleri ve kappa istatistiği ile sınırlandırılmıştır.

1.1. Kapsam Geçerlik Oranları ve İndeksi

Kapsam geçerlik oranı (KGO), geliştirilecek ölçekteki maddelerin kapsam geçerliğine ilişkin uzman görüşlerinin niceliklendirilmesinde kullanılan bir yaklaşımdır. Kapsam geçerlik indeksi (KGİ) ise kapsam geçerlik oranları gibi madde bazında değil tüm ölçek üzerinden elde edilen kapsam geçerlik düzeyini göstermektedir. Bir ölçek geliştirme çalışmasında konu alanı uzmanı, ilgili maddenin ölçülecek yapıyı temsil edip etmediğine ilişkin I) "tamamen temsil ediyor" (tamamen uygun), II) "oldukça temsil ediyor" (oldukça uygun), III) "biraz temsil ediyor" (biraz uygun) ve IV) "hiç temsil etmiyor" (uygun değil) şeklinde görüş bildirdiği varsayılsın. Davis (1992) durum I ve II ile görüş bildiren uzman sayısının toplam uzman sayısına oranını kapsam geçerlik oranı olarak nitelendirmiş ve bu oranın 0.80 ve daha yüksek olması halinde ilgili maddenin kapsam geçerliğinin yüksek olduğunu ifade etmiştir. Davis'in bu basit yaklaşımının yanı sıra; uzman görüşlerinin niceliklendirilmesi üzerine çeşitli çalışmalar söz konusudur (bkz: James vd., 1984; Lawshe, 1975; Lindell vd., 1999; Tinsley & Weiss, 1975). Bu çalışmalardan (hesaplama ve raporlama kolaylığı nedeniyle) en çok ön plana çıkan kapsam geçerlik ölçüsü Lawshe (1975) tarafından geliştirilen kapsam geçerlik oranı (madde bazında) ve kapsam geçerlik indeksidir (ölçek bazında). Lawshe'nin yaklaşımda geliştirilmesi amaçlanan ölçme aracındaki maddelerin ölçülmek istenilen özelliği ölçüp ölçmediğine ilişkin uzman görüşleri (her bir madde için) "gerekli/önemli", "yararlı ama önemli değil" ve "gereksiz" şeklinde ölçeklendirilmektedir. Elde edilen uzman görüşleri;

$$\text{KGOi} = \frac{N_G - \frac{N}{2}}{\frac{N}{2}} \qquad (1)$$

Burada KGOi i. maddenin kapsam geçerlik oranını göstermek üzere, NG i.madde için "gerekli" diyen uzman sayısını, N ise toplam uzman sayısını göstermektedir. Buna göre, KGOi, -1.0 ile 1.0 arasında değer alır. Araştırmaya katılan tüm uzmanlar ilgili maddeye "gerekli" şeklinde görüş bildirirse KGOi=1.0 ya da tüm uzmanlar "gerekli" dışındaki seçeneklere yönelirse KGOi=-1.0 değerini almaktadır. KGOi'nin pozitif ara değerleri için 0.05 anlamlılık düzeyinde (madde seçiminde kullanılmak üzere) minimum/kritik değerler Tablo 1'de verilmiştir (Lawshe, 1975).

Tablo 1: KGO için madde seçimine yönelik kritik değerler

Uzman Sayısı	Minimum Değer	Uzman Sayısı	Minimum Değer
5	0.99	13	0.54
6	0.99	14	0.51
7	0.99	15	0.49
8	0.75	20	0.42
9	0.78	25	0.37
10	0.62	30	0.33
11	0.59	35	0.31
12	0.56	40	0.29

Tablo 1 incelendiğinde; örneğin herhangi bir madde için 15 uzmanın görüşlerine dayanarak oluşturulmuş KGOi değeri 0.49'dan büyük ise bu durumda ilgili maddenin aday ölçme aracında yer alması yönünde karar verilir. Bununla birlikte; Wilson, Pan ve Schumsky (2012) yaptıkları bir çalışmada Lawshe tarafından önerilen ve Tablo 1 ile verilen kritik değerleri yeniden düzenlemişlerdir. Wilson, Pan ve Schumsky (2012) tarafından önerilen yeni kritik değerlere ilişkin çalışma bu araştırmanın raporlaştırılmasından sonra yayınlandığı için bu çalışmanın kapsamı dışında bırakılmıştır.

KGOi maddelerin ölçekte olması ya da olmamasına ilişkin kapsam geçerliğine dayalı bir madde istatistiğidir. Kapsam geçerlik indeksi (KGİ) ise ölçekte yer almasına karar verilen maddelerin KGOi ortalamalarından elde edilen bir test istatistiğidir (Lawshe, 1975). KGO'ları Şekil 1'de verilen ve ölçek geliştirme sürecinde yer alan tasarım aşamasının bir işlemidir ve madde havuzunun

yapılandırılmasında önemli bir yere sahiptir. İster madde bazında (KGO) ve isterse test bazında (KGİ) elde edilen bu niceliksel uzman görüşleri (kapsam geçerlik değerleri) özellikle yapı geçerliği sürecinin de bir ön koşuludur (Lawshe, 1985).

1.2. Kappa İstatistiği

1960 yılında Cohen, puanlayıcılar arasındaki uyum istatistiği olarak kappa istatistiğini geliştirmiştir. Cohen (1960) kappa istatistiğini puanlayıcılar arasındaki şans uyumu (chance aggrement, Pc) ile gözlenen uyum (observed aggrement, Po) arasındaki bağıntı üzerine yapılandırmıştır.

$$\text{Kappa}= \frac{P_o - P_C}{1 - P_C} \qquad (2)$$

Cohen'in bu katsayısı iki seçenekli durumlar (nominal) ve iki uzmanın uyumuna yöneliktir. Fleiss (1971) kappa istatistiğini daha fazla uzmanı kapsayacak şekilde geliştirmiştir. Ancak kappa istatistiği madde bazında kapsam geçerliği yerine ölçek bazında değerler üretmektedir. Polit, Beck ve Owen (2007) kappa istatistiğini madde bazında kapsam geçerliğine uygun hale getirmişlerdir.

$$\text{Kappa} = \frac{(\frac{N_G}{N}) - P_C}{1 - P_C} \qquad (3)$$

$$\text{PC} = \left[\frac{N!}{N_G!(N - N_G)!}\right].\left[\frac{1}{2}\right]^N \qquad (4)$$

Fleiss (1981), kappa değeri [0,60≤kappa≤0.74] arasında olduğunda uzman görüş uyumunun, ilgili madde için "iyi"; [kappa≥0,75] olduğunda ise uzman görüşleri arasındaki uyumun "mükemmel" olarak nitelendirilebileceğini ifade etmiştir.

Bu çalışmada, modifiye edilmiş kappa istatistiği ile kapsam geçerlik oranlarının bir ölçek geliştirme çalışmasındaki tutarlıkları ele alınarak incelenmiştir.

2. YÖNTEM

Bu araştırmanın uygulama bölümünde 10 maddelik bir ölçek geliştirme çalışmasının tasarım aşamasında maddelere ilişkin a) önce 5 uzmanın görüşü alınarak bunlar üzerinden kapsam geçerlik değerleri (Eşitlik 1 ve Eşitlik 3) hesaplanmış, b) sonra da farklı 5 uzmanın görüşü daha alınarak tekrar kapsam geçerlikleri bulunmuştur. Daha sonra kapsam geçerlik değerlerine göre olumsuz görünen maddeler ölçekten çıkartılmayarak ölçek geliştirme sürecinin uygulama aşamasına geçilmiştir. Burada aday ölçme aracı 132 bireye uygulanmış ve elde edilen sonuçlara bağlı olarak faktöriyel geçerlik değerlerine ulaşılmıştır. Bu değerlerden madde ayıredicilik indeksine karşılık gelen madde yükleri (λ) ve madde güvenirliği anlamına da gelen ortalama açıklanan varyans (AVE-Average variance extracted) değeri hesaplanmıştır.

$$\text{AVEi} = \frac{\lambda i \lambda i}{\lambda i \lambda i + \theta i} \qquad (5)$$

Uygulama bölümünde öncelikle aday ölçme aracında yer alması düşünülen her bir madde için KGO değerleri hesaplanmış ve sonra Tablo 1'de verilen kritik değerleri ile karşılaştırılarak 0.05

düzeyinde istatistiksel olarak anlamlı olanlar belirlenmiştir. Daha sonra kappa istatistikleri hesaplanarak Fleiss (1981) tarafından belirlenen aralıklara göre "iyi uyum" ve "mükemmel uyum" kapsamındaki maddeler belirlenmiştir.

3. UYGULAMA

Tablo 2'de 10 maddeye ilişkin 5 uzman görüşleri verilmiştir. Örneğin 1. maddeye 5 uzmandan 4'ü maddeye "ilgili madde ölçekte mutlaka olmalı" anlamına gelen "gerekli" şeklinde görüş bildirirken 1 tanesi ise maddenin ilgili boyutu ölçmede çok da önemli olmadığını ifade etmiştir. Bu maddenin KGO_1 değeri [(4-2.5)/2.5]=0.60 olarak hesaplanmıştır. Benzer şekilde PC değeri 0.16 ve kappa değeri ise 0.76 olarak hesaplanmıştır. Faktör analizi sonucu faktör yükleri (λ) ve AVE değerleri Tablo 2'de rapor edilmiştir.

Tablo 2: 5 uzmana ilişkin geçerlik parametre değerleri

Madde	Gerekli	Yetersiz	Gereksiz	KGO	PC	Kappa	λ	AVE
1	**4**	**1**	**0**	**0,60**	**0,16**	**0,76****	**0,82****	**0,75**
2	**3**	**1**	**1**	**0,20**	**0,31**	**0,42**	**0,69****	**0,62**
3	**5**	**0**	**0**	**1,00+**	**0,03**	**1,00****	**0,91****	**0,86**
4	**4**	**1**	**0**	**0,60**	**0,16**	**0,76****	**0,87****	**0,80**
5	**4**	**0**	**1**	**0,60**	**0,16**	**0,76****	**0,82****	**0,78**
6	**5**	**0**	**0**	**1,00+**	**0,03**	**1,00****	**0,89****	**0,83**
7	**3**	**2**	**0**	**0,20**	**0,31**	**0,42**	**0,58***	**0,38**
8	**5**	**0**	**0**	**1,00+**	**0,03**	**1,00****	**0,92****	**0,88**
9	**4**	**1**	**0**	**0,60**	**0,16**	**0,76****	**0,78****	**0,77**
10	**5**	**0**	**0**	**1,00+**	**0,03**	**1,00****	**0,96****	**0,88**

Tablo 2'de görüldüğü gibi 5 uzmanın görüşleri doğrultusunda KGO'na göre 0.05 anlamlılık düzeyinde sadece 3., 6., 8. ve 10. maddeler kapsam geçerliğinde olumlu maddeler olarak belirlenirken; kappa istatistiklerine göre 1., 3., 4., 5., 6., 8., 9. ve 10. maddelerde uzmanlar arasında olumlu yönde mükemmel bir uyumdan bahsedilebilir. Uygulama parametrelerinden faktör yüklerine ve AVE değerlerine bakıldığında, özünde tüm maddelerin faktöriyel geçerliğe yönelik istatistiksel olarak anlamlı katkı yaptığı görülmektedir. Buna ilişkin sınamalarda 0.01 (**) ve 0.05 (*) anlamlılık düzeyinde t testi ile test edilmiştir.

Tablo 2'de verilen madde bazındaki kapsam geçerlik ölçülerine ek olarak, test bazında ise kapsam geçerlik indeksi (KGİ) 0.68 olarak hesaplanmıştır. Fleiss'in (1971) çok puanlayıcılı kappa istatistik (multi-rater kappa statistics) değerinden ulaşılan test bazındaki kappa istatistiği ise 0.78 olarak elde edilmiştir. Benzer şekilde ölçeğin uygulanmasından sonra hesaplanan yapısal güvenirlik (construct reliability) değeri ise 0.77 olarak bulunmuştur.

Tablo 3'te ise 10 konu alanı uzmanın görüşleri ve bu görüşlere ilişkin kapsam geçerlik istatistikleri ile uygulama sonucu elde edilen istatistiklere yer verilmiştir. Sonuçlar incelendiğinde uzman sayısının 5'ten 10'a yükselmesiyle 5. ve 9. maddelere ilişkin KGO değerleri anlamlı hale gelmiştir. Ancak 10 maddenin 4 tanesinin Tablo 1'de verilen kritik değerlere yaklaşmasına karşın istatistiksel olarak anlamlı bulunmamıştır. Kappa istatistiği için ise; 10 uzman görüşüne dayalı değerlerde tüm maddeler olumlu uyum içinde bulunmuş, bunlardan sadece 7. madde "iyi uyum" konumuna yükselmiştir. Bu değerler ile faktör yük değerleri birlikte incelendiğinde kappa istatistiklerinin büyük bir tutarlılık içinde olduğu görülebilir.

Tablo 3: 10 uzmana ilişkin geçerlik parametre değerleri

Madde	Gerekli	Yetersiz	Gereksiz	KGO	PC	Kappa	λ	AVE
1	**8**	**2**	**0**	**0,60**	**0,04**	**0,79****	**0,82****	**0,75**
2	**8**	**1**	**1**	**0,60**	**0,04**	**0,79****	**0,69****	**0,62**
3	**9**	**1**	**0**	**0,80+**	**0,01**	**0,90****	**0,91****	**0,86**
4	**8**	**1**	**1**	**0,60**	**0,04**	**0,79****	**0,87****	**0,80**
5	**9**	**0**	**1**	**0,80+**	**0,01**	**0,90****	**0,82****	**0,78**
6	**10**	**0**	**0**	**1,00+**	**0,00**	**1,00****	**0,89****	**0,83**
7	**7**	**2**	**1**	**0,40**	**0,12**	**0,66***	**0,58***	**0,38**
8	**9**	**1**	**0**	**0,80+**	**0,01**	**0,90****	**0,92****	**0,88**
9	**8**	**1**	**1**	**0,60+**	**0,04**	**0,79****	**0,78****	**0,77**
10	**10**	**0**	**0**	**1,00+**	**0,00**	**1,00****	**0,96****	**0,88**

Tablo 3'te verilen değere ilişkin test bazında kapsam geçerlik indeksi 0.72, kappa istatistiği 0.81 ve uygulama sonucu elde edilen yapısal güvenirliğin önceki verilen değerinin 0.77 olduğu görülürse, uzman sayısının artmasıyla KGİ değerinin güvenirlik değerine yaklaştığı, kappa istatistiğinin de yapısal güvenirlik değerini aştığı görülebilir.

4. SONUÇ ve ÖNERİLER

Ölçek geliştirme sürecinde kapsam geçerliği yapı geçerliği için bir ön koşuldur. Aynı zamanda, bir ölçek geliştirme sürecinin pilot çalışması –ön uygulama- olanağı olmadığı durumlarda kapsam geçerliği tek başına da kullanılabilmektedir. Ancak kapsam geçerliklerinin (uzman görüşlerine dayalı olduğunda dolayı) raporlama ve iletişim güçlüğü nedeniyle nicel ifadelere dönüştürülmesi gerekmektedir. Bu konuda nitel yapıdaki uzman görüşlerini nicel ifadelere dönüştüren katsayıların başında Lawshe (1975) tarafından önerilen kapsam geçerlik oranı ve kapsam geçerlik indeksi gelmektedir. Bu katsayının bu denli popüler olmasının nedenlerinin başında basit, anlaşılabilir ve kolay hesaplanabilir yapıda olması gelmektedir. Ancak bu araştırma göstermiştir ki; kapsam geçerlik oranları, az sayıdaki uzmanlarla yapılan çalışmalarda kappa istatistiğine göre daha düşük değerler üretmekte, ayrıca istatistiksel olarak da maddelerin ölçekte yer alıp almamasına ilişkin kararlar bakımından daha az tutarlılık göstermektedir. Burada tartışılması gereken bir diğer konu da kapsam geçerlik oranlarının sağlam istatistik olup olmamasından daha çok maddelerin ölçekte kalıp kalmamasına yönelik kritik değerlerdir. Nicel değerler olarak kapsam geçerlik oranları ve/veya indekslerinin uzman sayısına bağlı olarak artış göstermesine karşın istatistiksel anlamlılık olarak bu niceliksel artışların uygun bulunmaması Tablo 1'deki minimum/kritik değerleri tartışılır hale getirmektedir. Zira son zamanlarda bu değerlerin yeniden yapılandırılması üzerine çalışmalar devam etmektedir (Wilson, Pan & Schumsky; 2012).

Diğer taraftan, madde bazındaki kappa istatistiği ise kapsam geçerlik oranlarına göre daha tutarlı sonuçlar üretmektedir. Bu sonuçlara göre araştırmacılara uzman sayılarının az olduğu durumlarda kapsam geçerlik oranları yerine kappa istatistiği önerilebilir. Bu öneri, araştırma bulgularında yer alan kappa değerleri, kappa değerlerinin değerlendirilmesi ve aynı zamanda uygulama aşamasındaki istatistikler ile tutarlılığına dayalı olarak yapılmaktadır. Örneğin; madde güvenirlik değerlerinin (AVE) yapı geçerliği için 0.5 değerinden yüksek olması önerilir (Fornell & Larcker; 1981). Tablo 3 incelendiğinde yalnızca 7. maddenin AVE değeri (AVE7=0.38) 0.5'ten küçük bulunmuştur. Bu maddenin kappa değerine göre yapılan değerlendirmede ise (geri kalan tüm maddeler üzerinden "mükemmel" bir uyum söz konusu iken) bu maddede sadece "iyi" uyumdan söz etmek olanaklıdır (0.66*).

KAYNAKLAR

Anastasi,A. (1988). *Psychological Testing*. (5th edition). New York: MacMillan Pub. Co. Inc.

Carmines, E.G., Zeller, R.A. (1979). *Reliability and Validity Assessment*, Quantitative Applications in the Social Sciences, Series No. 07–017. Sage Publications, Beverly Hills.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20 (1), 37–46.

Davis, L.L. (1992). Instrument review: Getting the most from a panel of experts. *Applied Nursing Research*, 5, 194–197.

DeVellis, R. F. (1991). *Scale development:Theory and applications*. Applied Social Research Methods Series, Volume 26. Newbury Park, CA: Sage Publications.

Fleiss, J. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76, 378–382.

Fleiss, J. (1981). *Statistical methods for rates and proportions* (2nd ed.). New York: John Wiley.

Fornell, C., & Larcker, D. (1981). Evaluating Structural Equation Models with Unobservable Variable and Measurement Error. *Journal of Marketing Research*, 18, 39-50.

James, L. R., Demaree, R. G., & Wolf, G. (1984). Estimating within-group interrater reliability with and without response bias. *Journal of Applied Psychology*, 69, 85–98.

James, L. R., Demaree, R. G., & Wolf, G. (1993). Rwg: An assessment of within-group interrater agreement. *Journal of* Applied Psychology, 78, 306–309.

Kerlinger, F.N. (1986). *Foundations of Behavioral Research*, 3rd edn. Holt, Rinehart and Winston, New York.

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology*, 28, 563–575.

Lawshe, C. H. (1985). Inferences from personnel tests and their validity. *Journal of Applied Psychology*, 70(1), 237-238.

Lindell, M. K., Brandt, C. J., & Whitney, D. J. (1999). A revised index of interrater agreement for multiitem ratings of a single target. *Applied Psychological Measurement*, 23, 127–135.

Nunnally, J.C. (1978). *Psychometric Theory*, 2nd edn. McGraw-Hill, New York.

O'Leary-Kelly, S.W. & Vokurka, R.J. (1998). The empirical assessment of construct validity. *Journal of Operations Management*, 16: 387-405.

Pedhazur, E.J., & Schmelkin, L.P. (1991). *Measurement, Design, and Analysis: An Integrated Approach*. Lawrence Erlbaum Associates, Publishers, Hillsdale, NJ.

Polit, D.F., Beck, C., & Owen, S.V. (2007). Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. *Research in Nursing & Health*, 30 (4), 459–467.

Tinsley, H. E. A., & Weiss, D. J. (1975). Interrater reliability and agreement of subjective judgments. *Journal of Counseling Psychology*, 22, 358–376.

Turner, S. (1979). The concept of face validity. *Quality and Quantity*, 13: 85-90.

Wilson, F. R., Pan, W., & Schumsky, D. A. (2012). Recalculation of the Critical Values for Lawshe's Content Validity Ratio. *Measurement and Evaluation in Counseling and Development*, DOI: 10.1177/0748175612440286.

Extended Abstract

It can be stated that there are two phases in the scale development process. The first phase is designing the scale; and the other phase is pre-application. There are different validity concepts specific to each phase. Particularly, while content and face validities have leading role in designing scale, the construct and criterion validities have come to the forefront in pre-application phase. In fact, content validity is prerequisite for construct validity. Content validity is often established through qualitative expert reviews. Content validity information often derives from reviews to be undertaken by subject matter experts as qualitative data. However, there are some studies about transforming these qualitative data into quantitative form; because it is difficult to report qualitative reviews of experts. Some of these studies are Kappa Statistics developed by Cohen (1960), T index of Tinsley–Weiss (Tinsley & Weiss, 1975), rWG and rWG(J) indexes proposed by James, Demaree ve Wolf (1993) and modified rWG(J) proposed by Lindell, Brandt and Whitney (1999). While some of them deal with the ratio of experts saying that the item is appropriate for content validity, and some are established on agreement of experts. This study was limited with Content Validity Index/ratio(CVI/CVR) and kappa statistics.

CVRi is an item statistic based on content validity related to whether the item must be in the scale or not. On the other hand Content Validity Index (CVI) is a test statistic obtained by the mean of items' CVR decided to put in the scale. (Lawshe, 1975).

Kappa is a statistic based on interrater agreement. It was modified for content validity on the basis of item by Polit, Beck and Owen (2007) based on the studies of Cohen (1960) and Fleiss (1971). They also stated that kappa value can be interpreted according to the classification of Fleiss (1981): ($60 \leq kappa \leq 0.74$: expert agreement for related item is "good"; $kappa \geq 0,75$: expert agreement for related item is "perfect").

In this research, consistency of modified kappa statistic and content validity ratio in developing scale process was examined. In the application phase of research, in the design phase for developing scale with 10 item a) first 5 experts' reviews were taken related with each item, and the value of content validity was calculated, b) next 5 different experts' reviews were taken too and the value of content validity was calculated again. The application phase of developing scale was begun without none of the items which seemed inappropriate for the content validity values was deleted. Draft scale was implemented to 132 people and then the data factor validity values (factor loading and AVE-Average variance extracted) were calculated.

Table 3: Validity values based on 5 experts' reviews

Item	Essential	Inefficient	Unneccessary	KGO	PC	Kappa	λ	AVE
1	**4**	**1**	**0**	**0,60**	**0,16**	**0,76****	**0,82****	**0,75**
2	**3**	**1**	**1**	**0,20**	**0,31**	**0,42**	**0,69****	**0,62**
3	**5**	**0**	**0**	**1,00+**	**0,03**	**1,00****	**0,91****	**0,86**
4	**4**	**1**	**0**	**0,60**	**0,16**	**0,76****	**0,87****	**0,80**
5	**4**	**0**	**1**	**0,60**	**0,16**	**0,76****	**0,82****	**0,78**
6	**5**	**0**	**0**	**1,00+**	**0,03**	**1,00****	**0,89****	**0,83**
7	**3**	**2**	**0**	**0,20**	**0,31**	**0,42**	**0,58***	**0,38**
8	**5**	**0**	**0**	**1,00+**	**0,03**	**1,00****	**0,92****	**0,88**
9	**4**	**1**	**0**	**0,60**	**0,16**	**0,76****	**0,78****	**0,77**
10	**5**	**0**	**0**	**1,00+**	**0,03**	**1,00****	**0,96****	**0,88**

Table 4: Validity values based on 10 experts' reviews

Item	Essential	Inefficient	Unneccessary	KGO	PC	Kappa	λ	AVE
1	**8**	**2**	**0**	**0,60**	**0,04**	**0,79****	**0,82****	**0,75**
2	**8**	**1**	**1**	**0,60**	**0,04**	**0,79****	**0,69****	**0,62**
3	**9**	**1**	**0**	**0,80+**	**0,01**	**0,90****	**0,91****	**0,86**
4	**8**	**1**	**1**	**0,60**	**0,04**	**0,79****	**0,87****	**0,80**
5	**9**	**0**	**1**	**0,80+**	**0,01**	**0,90****	**0,82****	**0,78**
6	**10**	**0**	**0**	**1,00+**	**0,00**	**1,00****	**0,89****	**0,83**
7	**7**	**2**	**1**	**0,40**	**0,12**	**0,66***	**0,58***	**0,38**
8	**9**	**1**	**0**	**0,80+**	**0,01**	**0,90****	**0,92****	**0,88**
9	**8**	**1**	**1**	**0,60+**	**0,04**	**0,79****	**0,78****	**0,77**
10	**10**	**0**	**0**	**1,00+**	**0,00**	**1,00****	**0,96****	**0,88**

According to the result, content validity ratio, makes lower value with regard to kappa statistic in the studies which were done with experts in small numbers; in addition it shows less consistency in deciding whether the item must be in the scale or not. Another topic to be discussed here is the minimum value for whether the item must be in the scale or not rather than content validity ratios. Likewise the studies on modifying these minimum values have been continuing recently (Wilson, Pan & Schumsky; 2012).

On the other hand, the item based kappa statistic is producing more consistent results than the content validity ratio. According to these results, instead of content validity ratio, kappa statistic can be suggested to the researchers in case of having experts in small numbers.