The Korean Journal of Applied Statistics (2012) DOI: http://dx.doi.org/10.5351/KJAS.2012.25.5.719
25(5), 719–731

A Study on Comparison of Generalized Kappa Statistics in Agreement Analysis

Min Seon Kim[1] · Ki Jun Song[2] · Chung Mo Nam[3] · Inkyung Jung[4]

[1]Department of Biostatistics, Yonsei University College of Medicine
[2]Department of Biostatistics, Yonsei University College of Medicine
[3]Department of Biostatistics, Yonsei University College of Medicine
[4]Department of Biostatistics, Yonsei University College of Medicine

(Received June 4, 2012; Revised September 17, 2012; Accepted September 19, 2012)

Abstract

Agreement analysis is conducted to assess reliability among rating results performed repeatedly on the same subjects by one or more raters. The kappa statistic is commonly used when rating scales are categorical. The simple and weighted kappa statistics are used to measure the degree of agreement between two raters, and the generalized kappa statistics to measure the degree of agreement among more than two raters. In this paper, we compare the performance of four different generalized kappa statistics proposed by Fleiss (1971), Conger (1980), Randolph (2005), and Gwet (2008a). We also examine how sensitive each of four generalized kappa statistics can be to the marginal probability distribution as to whether marginal balancedness and/or homogeneity hold or not. The performance of the four methods is compared in terms of the relative bias and coverage rate through simulation studies in various scenarios with different numbers of raters, subjects, and categories. A real data example is also presented to illustrate the four methods.

Keywords: Agreement, generalized kappa, marginal probability distribution.

1. 서론

동일한 측정 대상들에 대해 평가자(rater)들이 평가한 결과가 일치하는 정도를 일치도(agreement)로 나타낸다. 카파통계량(kappa statistic)은 측정한 결과가 범주형 자료일 때 일치도의 척도로 자주 쓰인다. 단순 카파통계량(simple kappa; Cohen, 1960)이나 가중 카파통계량(weighted kappa; Cohen, 1968)은 평가자가 둘인 경우 사용되고, 평가자가 세 명 이상인 경우에는 일반화 카파통계량(generalized kappa) κ가 사용된다.

본 연구에서는 일반화 카파통계량으로 제안된 여러 방법들이 주변확률분포(marginal probability distribution)를 변화시키면서 어느 정도 민감하게 반응하는지, 그 원인은 무엇인지 알아본다. 또한 평가자 수, 표본수, 범주수가 변화함에 따른 일반화 카파통계량 값을 비교, 평가하고자 한다. Scott의 π (Scott, 1955)의 개념을 확장하여 제안한 Fleiss의 방법 (Fleiss, 1971), Cohen (Cohen, 1960)의 카파통계량

[4]Corresponding author: Assistant Professor, Department of Biostatistics, Yonsei University College of Medicine, 50 Yonsei-ro, Seodaemun-gu, Seoul 120-752, Korea. Email: ijung@yuhs.ac

을 확장하여 Fleiss의 일반화 카파통계량을 보정한 Conger의 방법 (Conger, 1980), 주변분포에 영향을 받지 않는 Randolph의 방법 (Randolph, 2005), 평가자 내 변동을 나타내는 불확실성 계수를 이용한 Gwet의 방법 (Gwet, 2008a)인 AC1 통계량을 고려한다. 위의 방법들의 장, 단점을 살펴보기 위해 프로그램(SAS 9.2)을 이용하여 모의실험을 한다. 주변확률이 변화함에 따라 각 방법으로 구한 일반화 카파통계량이 어떤 양상으로 반응하는지 알아보고, 민감하게 반응하는 문제점의 원인을 균형적 주변분포와 주변동질성 여부를 중심으로 찾아본다.

2절에서는 일치도의 소개와 단순 카파통계량의 두 가지 문제점, 일반화 카파통계량에 대한 개념과 함께 일반화 카파통계량 방법에 관한 이론적 배경을 언급한다. 3절에서는 주변확률을 변화시키면서 2.3절에서 소개한 네 가지 방법을 다양한 수준에서 비교, 확인한다. 4절에서는 모의실험을 통하여 평가자수, 평가대상자수, 범주수를 다양하게 변화시키며 네 가지 방법의 추정값의 정확성을 비교한다. 5절에서는 아동 환자의 수신증 자료에 네 가지 방법을 적용한 결과를 비교한다. 마지막으로 6절에서 고찰 및 결론을 제시한다.

2. 이론적 배경

2.1. 일치도

일치도란 한 표본을 여러 번 반복 측정한 결과가 서로 어느 정도 일치하는가를 알아보는 신뢰도 평가의 척도로, 한 명의 평가자가 한 표본을 반복 측정, 혹은 여러 명의 평가자가 한 표본을 평가할 때 일치하는 정도이다. 연속형 자료에서는 신뢰도를 나타내기 위한 상관 계수로 급내상관계수(Intra-class correlation coefficient)가 정의되었고, 범주형 자료에서는 Cohen (1960)이 처음으로 두 명의 평가자 간의 일치도를 제시하여 카파통계량을 정의하였다. 그 후 평가자가 여러 명인 경우 (Fleiss, 1971) 등 다양한 카파통계량이 제안되었다.

2.2. 단순 카파통계량의 두 가지 역설

단순 카파통계량은 고정된 한 쌍의 평가자가 표본을 명목형의 범주로 분류하였을 때의 일치도를 측정할 때 쓰인다. 카파통계량의 표현식은 다음과 같다. 평가자들이 대상자들을 우연히 같은 범주로 분류하는 경우가 있으므로, 그 확률을 보정한 일치도를 사용한다.

$$\kappa = \frac{P_a - P_e}{1 - P_e}, \tag{2.1}$$

여기서 P_a는 관찰된 일치비율로 두 평가자가 같은 범주로 분류한 평가대상자들의 비율이다. P_e는 두 평가자가 독립이라는 가정 하에 우연에 의해 기대되는 일치비율이며, κ값은 관찰된 일치비율과 완벽하게 일치할 경우의 비율인 '1'에서 각각 우연에 의한 일치비율을 뺀 값의 비로 정의되고 1에 가까울수록 일치도가 높다고 할 수 있다.

P_a의 크기가 동일하다면 P_e가 작을수록 κ는 커진다. 따라서 P_a가 1에 가까운 값이라고 해도, P_e가 크다면 κ는 매우 작은 값을 가진다. P_e가 주변분포에 의존한다면 그 분포가 다르기 때문에 나타난 P_e의 차이가 κ값에 영향을 미치게 된다. κ의 이러한 문제점을 Feinstein과 Cicchetti (1990)는 평가자수와 범주수가 2인 상황으로 한정하여 원인을 크게 두 가지로 설명한다.

Table 2.1에서 각 평가자가 평가대상자들을 두 범주에 할당한 주변확률이 각각 0.5에 가까운 경우($(a+b)/n \approx 0.5$, $(a+c)/n \approx 0.5$)에는 주변분포가 균형적(balanced marginal distribution)이라고 하고, 주변확률이 0.5보다 상당히 크거나 작은 경우는 주변분포가 불균형적이라고 한다(unbalanced marginal distribution). 평가자 두 명의 각 범주에서의 주변분포가 같을 때, 즉 $(a+b) = (a+c)$일 경우, 이

Table 2.1. Binary rating results between two raters

평가자2	평가자1		총
	+	−	
+	a	b	$a+b$
−	c	d	$c+d$
총	$a+c$	$b+d$	n

Table 2.2. Data structure for calculating generalized kappa statistics

평가대상자	범주				평가자수
	1	2	⋯	q	
1	r_{11}	r_{12}	⋯	r_{1q}	r
2	r_{21}	r_{22}	⋯	r_{2q}	r
⋮	⋮	⋮	⋱	⋮	⋮
n	r_{n1}	r_{n2}	⋯	r_{nq}	r
총	r_{+1}	r_{+2}	⋯	r_{+q}	nr

를 주변동질성(marginal homogeneity)을 만족한다고 한다. κ의 두 가지 역설은 관찰된 일치비율 $P_a = (a+d)/n$가 동일하더라도, 주변분포가 균형적일 때가 그렇지 않을 때 보다 κ가 크고, 불균형 주변분포에서는 주변동질성을 만족하지 않는 경우가 만족하는 경우보다 κ가 크다는 것이다. 이로 인해 관찰된 일치비율이 큼에도 불구하고 κ가 많이 낮아질 수 있다는 것이 κ의 큰 문제점이 된다.

제 3절에서는 평가자와 범주가 3 이상인 경우로 확장시켜 Feinstein과 Cicchetti (1990)이 설명한 κ의 문제점이 2.3절의 네 가지 일반화 카파통계량 방법에서는 어떤 양상으로 작용하는지 알아보고자 한다.

2.3. 일반화 카파통계량

r명의 평가자가 n명의 평가대상자를 q개의 범주로 평가한다고 가정할 때, 일반화 카파통계량을 구하기 위해서는 자료 구조를 Table 2.2에서처럼 평가대상자와 범주에 관하여 나타내어야 한다. r_{ij}는 i번째 평가대상자를 j번째 범주로 평가한 평가자수이다.

일반화 카파통계량의 표현식은 단순 카파통계량과 동일하게 식 (2.1)과 같다. 다만, P_a는 다음 식과 같이 표현할 수 있고, 우연에 의해 기대되는 일치비율 P_e의 정의에 따라 여러 가지 일치도가 정의된다.

$$P_a = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{q}\frac{r_{ij}(r_{ij}-1)}{r(r-1)} = \frac{1}{nr(r-1)}\left(\sum_{i=1}^{n}\sum_{j=1}^{q} r_{ij}^2 - rn\right).$$

2.3.1. Fleiss의 방법 Scott의 π (Scott, 1955)의 개념을 확장시킨 형태로, 평가 결과의 범주는 순서를 고려하지 않는 이분형이나 명목형이고, 평가자는 서로 독립임을 가정한다. 우연에 의해 기대되는 일치비율을 $P_e = \sum_{j=1}^{q} p_j^2$로 표현한다. 여기서 $p_j = (1/n)\sum_{i=1}^{n}(r_{ij}/r)$로 평가를 하는 모든 경우의 수 nr에서 j번째 범주에 각 평가대상자를 분류한 평가자수의 비율이며, 위의 P_e로 보정한 Fleiss 방법의 κ 식은 다음과 같다.

$$\kappa = \frac{\frac{1}{nr(r-1)}\left(\sum_{i=1}^{n}\sum_{j=1}^{q} r_{ij}^2 - rn\right) - \sum_{j=1}^{q} p_j^2}{1-\sum_{j=1}^{q} p_j^2} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{q} r_{ij}^2 - rn\left(1+(r-1)\sum_{j=1}^{q} p_j^2\right)}{nr(r-1)\left(1-\sum_{j=1}^{q} p_j^2\right)}.$$

위 통계량의 분산은 평가자 사이에 일치도가 없다는 가정 하에 제안되었다. 평가대상자수 n이 크면 $p_1, \ldots, p_j$는 상수라고 볼 수 있기 때문에 분산식에서 $\sum_{j=1}^{q} r_{ij}^2$만 확률변수이고 평가대상자 간에 독립이므로 식을 정리하면 다음과 같다.

$$\mathrm{Var}(\kappa) = \frac{2}{nr(r-1)} \frac{\sum_{j=1}^{q} p_j^2 - (2r-3)\left(\sum_{j=1}^{q} p_j^2\right)^2 + 2(r-2)\sum_{j=1}^{q} p_j^3}{\left(1 - \sum_{j=1}^{q} p_j^2\right)^2}.$$

2.3.2. Conger의 방법 Cohen의 κ (Cohen, 1960)의 개념을 확장한 형태로 Fleiss의 exact kappa라고도 불린다. Conger의 방법에서 P_e는 아래와 같이 제시된다.

$$P_e = \sum_{j=1}^{q} p_j^2 - \sum_{j=1}^{q} \frac{s_j^2}{r}, \quad \text{where} \quad s_j^2 = \frac{1}{r-1}\sum_{k=1}^{r}(p_{jk} - \bar{p}_{j\cdot})^2.$$

위 식에서 $p_{jk} = n_{jk}/n$로 n_{jk}는 k번째 평가자가 j범주로 분류한 평가대상자수이며 s_j^2는 각 범주에서 평가자들의 주변확률에 대한 표본분산이다. Fleiss 방법의 P_e에서 평가자들의 변동 효과를 제거함으로써, 위의 식으로 κ를 표현해 보면 다음과 같다.

$$\kappa = \frac{\frac{1}{nr(r-1)}\left(\sum_{i=1}^{n}\sum_{j=1}^{q} r_{ij}^2 - rn\right) - \left(\sum_{j=1}^{q} p_j^2 - \sum_{j=1}^{q}\frac{s_j^2}{r}\right)}{1 - \left(\sum_{j=1}^{q} p_j^2 - \sum_{j=1}^{q}\frac{s_j^2}{r}\right)}.$$

Conger의 방법은 평가자수가 4이상인 경우에는 Fleiss의 방법으로 구한 값과 거의 비슷해지는 경향이 있다.

2.3.3. Randolph의 방법 2.2절에서 언급했듯이 관찰된 일치비율 P_a가 1에 가깝더라도 우연에 의해 기대되는 일치비율 P_e가 크다면 κ는 작아진다. Brennan과 Prediger (1981)는 평가자가 두 명인 경우, 2.2절에 설명한 두 가지 역설을 해결하고자 주변분포에 영향을 받지 않는 카파통계량을 제안하였다. 주변분포가 사전에 정의되지 않았을 경우, 각 평가자가 무작위로 평가대상자를 q개의 범주에 분류하였을 때, 주변 비율의 기대값은 각 범주마다 $1/q$이다. 그러므로 우연에 의해 기대되는 일치가 범주 k에서 일어났을 확률은 $(1/q) \times (1/q) = 1/q^2$이고, 각 범주의 확률의 합은 $q/q^2 = 1/q$이다. 따라서 Brennan과 Prediger (1981)는 P_e를 $1/q$로 정의한다.

Randolph의 논문은 Brennan과 Prediger의 통계량을 평가자가 세 명 이상인 경우에도 확장시켜 일반화 카파통계량을 제안한다.

$$\kappa = \frac{\frac{1}{nr(r-1)}\left(\sum_{i=1}^{n}\sum_{j=1}^{q} r_{ij}^2 - rn\right) - \frac{1}{q}}{1 - \frac{1}{q}}.$$

Fleiss와 Conger의 방법은 P_a가 같더라도 주변분포의 주변동질성과 균형성 여부에 따라 κ값이 크게 변화하는 반면, Randolph의 방법은 주변분포에 영향을 받지 않고 단지 범주의 수에 인해 변화한다.

2.3.4. Gwet의 방법 Randolph의 방법과 마찬가지로 Gwet (2008a)의 방법은 κ값이 P_e에 민감하지 않도록 보완된 통계량으로 다음과 같이 정의된다.

$$\kappa_{AC1} = \frac{\frac{1}{nr(r-1)}\left(\sum_{i=1}^{n}\sum_{j=1}^{q} r_{ij}^2 - rn\right) - \frac{1}{q-1}\sum_{j=1}^{q} p_j(1-p_j)}{1 - \frac{1}{q-1}\sum_{j=1}^{q} p_j(1-p_j)}.$$

Gwet의 방법에서는 분산식을 평가대상자의 변동만을 고려하였을 경우와 평가대상자와 평가자의 변동을 모두 고려하였을 경우로 구분하여 제안하였다 (Gwet, 2008b). 평가자가 CT나 MRI와 같이 정해져 있는 특정 기계일 때에는 평가자의 변동성을 고려하지 않아도 되는 반면, 의사나 간호사 집단에서 몇 명을 추출하여 평가자로 투입이 된다면 평가자의 변동성도 고려해야 할 것이다. 이 논문에서는 이런 경우를 분류하여 적용하기를 권하고 있다.

3. 주변확률 변화에 따른 비교

2.2절에서 언급했듯이 κ값이 주변분포에 민감하게 되면 실제로 일치도가 높아 보이는 자료라도 계산된 κ값은 그렇지 않은 경우가 종종 발생한다. 따라서 일반화 카파통계량도 평가자가 두 명인 경우와 마찬가지로 자료가 범주에 대해 균형적 주변분포, 주변동질성의 여부에 따라 κ값이 크게 영향을 받는지를 주변확률을 다양하게 변화시키면서 알아보고자 한다.

우선 균형적 주변분포를 만족할 경우, r명의 평가자를 두 명씩 묶었을 때 $\binom{r}{2}$가지 경우의 수 모두 주변동질성의 만족 여부에 따라서 P_e가 달라지는지 비교해보기 위해 평가자수와 범주수가 3이고 평가대상자수가 30명인 조건에서 확인해 보았다. 그리고 다양한 주변확률에서 일반화 카파통계량의 네 가지 방법의 변화하는 양상을 비교하기 위해서 평가자수는 세 명으로 고정하고 범주수를 2, 3, 4인 경우를 각각 적용하였고 범주수가 2, 3인 경우는 평가대상자를 30명, 4인 경우는 50명으로 자료를 생성하였다. 관찰된 일치비율 P_a는 모두 0.8로 같도록 데이터에 적용하여 P_e의 차이로 인한 κ값의 변화를 비교하였다.

균형적 주변분포(balanced marginal distribution)를 만족할 때, 세 명의 평가자를 두 명씩 묶은 세 가지 경우의 수 모두 주변동질성(marginal homogeneity)을 만족하는 경우와 만족하지 않는 경우의 자료를 생성하여 P_e를 비교하였다.

Table 3.1의 왼쪽 열은 평가자 세 명 모두 서로 주변동질성을 만족하는 경우이고, 오른쪽 열은 주변동질성이 만족하지 않는 경우의 자료이다. Table 3.1을 Table 2.2의 자료 구조 형태로 정리한다면 범주 별 평가대상자의 비율이 각각 1/3으로, 두 경우 모두 일반화 카파통계량의 네 가지 방법의 P_e값이 모두 같다. 따라서 일반화 카파통계량에서는 주변동질성 여부가 κ값에 영향을 미치지 않는 것을 알 수 있다.

하지만 주변확률이 불균형하게 변화할수록 각 방법의 κ값은 다양하게 변화하였다. 네 방법으로 구한 κ값의 비교는 Figure 3.1에 나타냈다. 균형적 주변분포를 따르는 경우부터 적어도 한 쌍에서 균형적 주변분포를 따르는 경우, 모든 범주에서 따르지 않는 경우 순으로 표현하였다. 여기서 균형적 주변분포란, Table 2.2에서 $r_{+1}/nr = r_{+2}/nr = \cdots = r_{+q}/nr = 1/q$인 경우이다.

Figure 3.1을 보면 각 범주 간의 주변확률의 차이가 클 경우, 즉 균형적 주변분포를 심하게 위배하는 경우 Fleiss와 Conger의 방법은 κ값이 매우 작아지고 거의 비슷한 값으로 증감하고 있다. 일반화 카파통계량에서 이 두 방법은 평가자수가 2인 경우와는 달리 주변동질성 여부에는 영향을 받지 않고 오로지 균형적 주변분포의 여부에 영향을 크게 받는다. 평가자의 수가 동일할 때 범주수가 늘어날수록 그 영향은 줄어든다. Figure 3.1에서 범주수가 2인 경우를 보면 불균형적 주변분포의 정도가 상당히 큰 (0.1,

Table 3.1. Data example when marginal homogeneity holds or not

평가자 1/2

	범주1	범주2	범주3	총
범주1	8	1	1	10
범주2	2	8	0	10
범주3	0	1	9	10
총	10	10	10	30

평가자 2/3

	범주1	범주2	범주3	총
범주1	8	2	0	10
범주2	1	7	2	10
범주3	1	1	8	10
총	10	10	10	30

평가자 1/3

	범주1	범주2	범주3	총
범주1	9	1	0	10
범주2	1	7	2	10
범주3	0	2	8	10
총	10	10	10	30

평가자 1/2

	범주1	범주2	범주3	총
범주1	9	3	1	13
범주2	2	6	1	9
범주3	0	0	8	8
총	11	9	10	30

평가자 2/3

	범주1	범주2	범주3	총
범주1	8	3	0	11
범주2	0	9	0	9
범주3	1	1	8	10
총	9	13	8	30

평가자 1/3

	범주1	범주2	범주3	총
범주1	8	4	1	13
범주2	0	9	0	9
범주3	1	0	7	8
총	9	13	8	30

0.9)의 경우에는 P_a는 0.8인데도 불구하고 Fleiss와 Conger의 방법은 κ값이 음수인 것을 볼 수 있다. 이 두 방법은 범주 간 주변확률이 균형성을 만족하는지 여부에 따라 값이 변동하였다.

Randolph의 방법은 P_e가 모든 경우의 수에서 범주수의 역수, 즉 $1/q$로 나눠주므로 κ도 항상 동일하다. 따라서 평가자수가 클수록 Randolph의 방법으로 구한 κ값은 P_a와 매우 비슷해질 것을 예상할 수 있다. Gwet의 방법은 Figure 3.1에서 보면 주변동질성이 만족하지 않을수록 만족할 때보다 오히려 κ값이 큰 양상을 보였다. 그 차이가 범주수가 2인 경우에는 큰 편이었지만 범주수가 늘어나면서 차이도 작아지고 Randolph의 방법의 값과 거의 같아진다. 각 범주의 주변확률이 같아질수록 즉 균형적 주변분포를 만족할수록 네 가지 방법 모두 같은 값으로 수렴함을 볼 수 있다.

4. 모의실험을 통한 비교

4.1. 모의실험 설계

2.3절에서 다룬 네 가지 방법을 비교하기 위하여 평가자의 변동성을 고려하지 않는 모집단과 고려하는 모집단을 생성하여 접근하였다. 평가대상자수 N은 500명, 평가자수 R은 각각 3, 5, 7, 9명이 되도록 4개의 모집단을 생성하였고, 평가대상자와 평가자의 변동성을 모두 고려한 모집단의 평가대상자수 N은 500명, 평가자수 R은 20명으로 하나의 모집단을 생성하였다. 범주수는 3과 5인 경우로 한정하였다. R명의 평가자들이 n명의 평가대상자에 대하여 각각 무작위로 조건에 따라 셋 혹은 다섯 범주에 할당하여 균형적 주변분포를 따르도록 설정하였다. 각각의 모집단에서 평가자 N명은 0.8의 확률로 같은 값을 평가하고 0.2의 확률로는 평가가 무작위로 이루어지도록 설계하였다.

위의 각 모집단에서 얻은 네 가지 방법의 일반화 카파통계량 κ는 모집단값으로 간주하고 아래의 조건에서 구한 κ값이 모집단값과 얼마나 가까운지 알아보고자 한다. 본 연구에서는 모집단 N에서 평가대상자

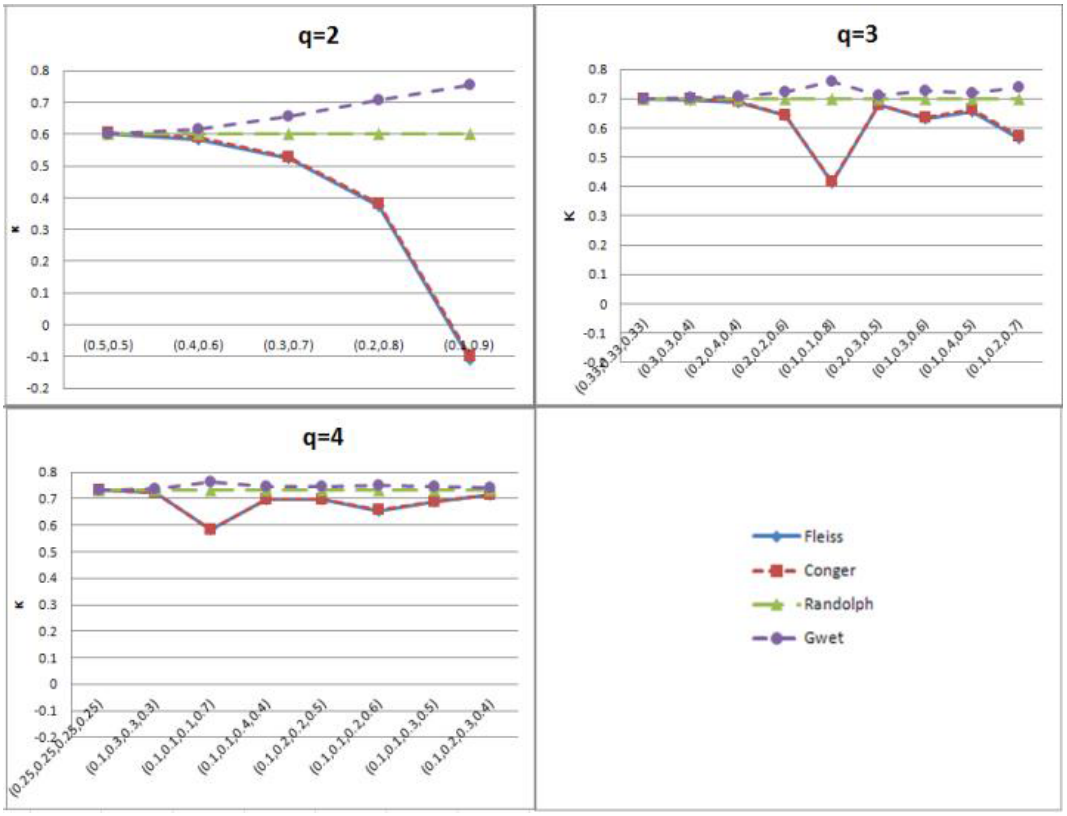

Figure 3.1. κ values according to change of marginal probabilities with $r = 3$ and $q = 2, 3, 4$

의 표본수 n을 각각 50, 100, 200으로 설정하고, 평가자의 변동성을 고려한 경우에는 평가자의 표본수 r이 각각 3, 5, 7, 9가 되도록 하였다. 모의실험을 수행하기 위해 위에서 설정한 각 경우의 모집단에서 다양한 조건의 무작위 표본을 각각 1000개씩 생성하였다.

평가대상자와 평가자의 표본수가 작을 때, Fleiss의 방법은 우연에 의한 일치 확률인 P_e가 1이 되어 일반화 카파통계량의 분모가 0이 될 수 있다. 이 경우를 보완하기 위하여 P_e가 1에 가까운 경우 이 값은 0.99999로 대체하였다.

위의 설정을 통해 표본에서 네 가지 방법으로 구한 추정된 일반화 카파통계량 값이 모집단값을 얼마나 정확하게 추정하는지에 대해서는 아래의 relative bias의 식을 통해 확인하였다.

$$\mathrm{RB}(\hat{\kappa}) = 100 \times \left(\frac{1}{1000} \sum_{s=1}^{1000} \hat{\kappa}_s - \kappa \right) / \kappa(\%)$$

κ는 모집단에서 구한 일반화 카파통계량 모집단값이고, $\hat{\kappa}_s$는 생성된 무작위 표본에서 구한 일반화 카파통계량 값이다. 또한 일반화 카파통계량의 κ값과 분산을 이용하여 95%의 신뢰구간 안에 모집단값이 얼마나 많이 포함되는지를 coverage rate으로 정의하여 계산하였다. 분산의 식이 없는 Conger와 Randolph의 방법은 잭나이프 방법 (Quenouille, 1949)을 통해 분산을 구하였다.

Table 4.1. Population κ values when ignoring the rater sampling variability

q	R	N	κ_F	κ_C	κ_R	κ_{AC1}
3	3	500	0.6738	0.6738	0.6740	0.6741
	5		0.6230	0.6230	0.6232	0.6233
	7		0.6314	0.6314	0.6317	0.6319
	9		0.6352	0.6352	0.6355	0.6357
5	3	500	0.6778	0.6779	0.6783	0.6785
	5		0.6384	0.6385	0.6388	0.6388
	7		0.6573	0.6574	0.6581	0.6583
	9		0.6337	0.6338	0.6348	0.6351

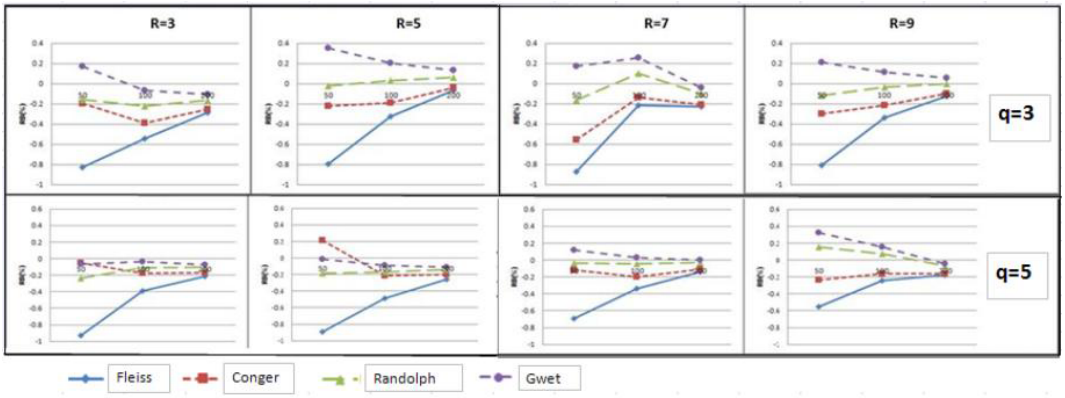

Figure 4.1. Relative bias (%) of the four generalized kappa statistics when ignoring the rater sampling variability

4.2. 결과

4.2.1. 평가자의 변동성을 고려하지 않는 경우 평가대상자수 N이 500, 평가자수 R이 각각 3, 5, 7, 9이고 범주수가 각각 3 또는 5인 모집단에서 표본의 평가대상자수 n을 각각 50, 100, 200으로 추출하였다. 표본의 평가자수는 모집단의 평가자수로 고정이며, 네 가지 방법의 카파통계량의 모집단값은 Table 4.1에, 표본의 평가대상자수에 따른 relative bias의 변화는 Figure 4.1에, coverage rate은 Table 4.2에 제시하였다.

전체적으로 Randolph와 Gwet의 방법이 Fleiss와 Conger의 방법에 비해 relative bias가 작게 나타났다. 특히 Fleiss의 방법은 평가대상자수가 50인 경우에는 다른 방법들에 비해 relative bias가 음수로 크게 나타났고 평가대상자수가 커지면서 그 편향의 크기는 큰 폭으로 작아지지만 다른 방법들에 비해서는 여전히 큰 값을 나타낸다. Fleiss와 Conger의 방법은 모든 조건에서 κ값을 과소추정하는 경향을 보였다. 네 방법 모두 평가대상자수가 커질수록 relative bias가 0에 가깝게 수렴하고 있다. Table 4.2에서 보듯이, Fleiss 방법을 제외한 나머지 방법들은 거의 모든 경우에 있어 coverage rate이 95%보다 약간 큰 값으로 얻어졌다. Fleiss의 방법은 coverage rate이 굉장히 낮게 나타났고 65%에서 95%까지 분포의 폭이 컸다.

4.2.2. 평가자의 변동성을 고려한 경우 평가대상자수 N은 500명, 모집단의 평가자수 R은 20명이고 범주수가 각각 3 또는 5인 모집단에서 평가대상자의 표본수 n은 50, 100, 200명으로 평가자의 표본수 r은 R에서 무작위로 3, 5, 7, 9명을 추출하였다. 모집단값은 Table 4.3에, 표본의 평가대상자수에 따른

Table 4.2. Coverage rates (%) of confidence intervals for the four generalized kappa statistics when ignoring the rater sampling variability

q	R	n	CR($\hat{\kappa}_F$)	CR($\hat{\kappa}_C$)	CR($\hat{\kappa}_R$)	CR($\hat{\kappa}_{AC1}$)
3	3	50	90.6	96.2	95.9	95.8
		100	94.0	96.5	96.4	96.5
		200	96.5	98.8	98.8	98.7
	5	50	83.1	96.4	96.2	96.1
		100	83.7	97.1	96.8	96.8
		200	88.0	98.6	98.6	98.7
	7	50	71.2	96.3	95.6	95.6
		100	74.1	97.6	97.4	97.5
		200	79.2	98.5	98.5	98.5
	9	50	65.5	95.9	95.5	95.9
		100	65.3	96.9	96.7	96.7
		200	73.1	98.3	98.1	98.1
5	3	50	85.5	95.4	95.1	95.2
		100	96.4	96.4	96.5	96.4
		200	99.2	99.2	99.3	99.3
	5	50	69.1	96.6	96.0	96.0
		100	71.6	97.0	96.8	96.9
		200	75.4	98.7	98.6	98.5
	7	50	62.1	94.8	95.2	94.8
		100	62.5	97.5	97.3	96.8
		200	67.8	98.7	98.7	98.7
	9	50	59.8	96.4	96.0	95.9
		100	58.6	96.1	95.4	95.4
		200	63.2	98.9	98.8	98.8

Table 4.3. Population κ values when considering the rater sampling variability

q	R	N	κ_F	κ_C	κ_R	κ_{AC1}
3	20	500	0.5698	0.5699	0.6307	0.6551
5			0.6166	0.6166	0.6167	0.6167

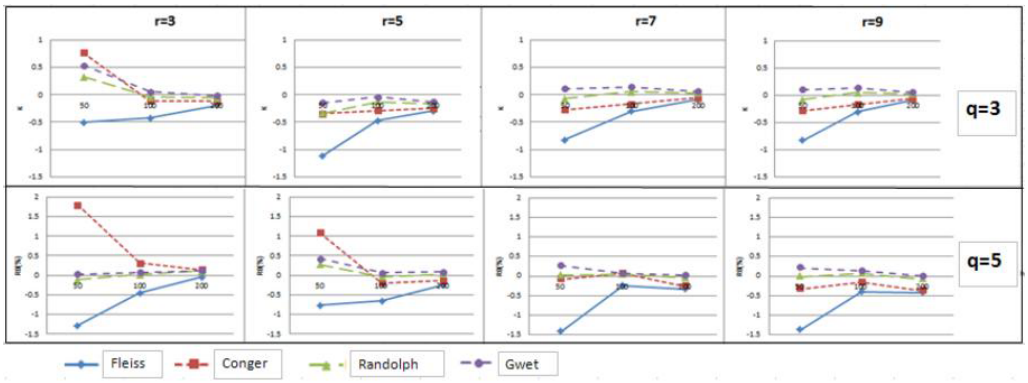

Figure 4.2. Relative bias (%) of the four generalized kappa statistics when considering the rater sampling variability

relative bias의 변화를 평가자수 별로 Figure 4.2에, coverage rate은 Table 4.4에 제시하였다.

결과는 평가자의 변동성을 고려하지 않는 경우 마찬가지로 Randolph와 Gwet의 방법이 Fleiss와 Conger의 방법에 비해 relative bias가 작게 나타났는데, Fleiss와 Conger의 방법은 평가자의 변동성을 고

Table 4.4. Coverage rates (%) of confidence intervals for the four generalized kappa statistics when considering the rater sampling variability

q	r	n	CR($\hat{\kappa}_F$)	CR($\hat{\kappa}_C$)	CR($\hat{\kappa}_R$)	CR($\hat{\kappa}_{AC1}$)
3	3	50	93.5	93.9	92.3	93.1
		100	94.0	95.4	94.6	95.4
		200	97.2	97.9	97.1	97.1
	5	50	91.0	95.5	95.7	97.9
		100	90.2	96.4	95.7	98.1
		200	93.7	98.3	96.9	98.5
	7	50	83.6	95.6	93.8	97.9
		100	88.2	97.1	96.0	98.7
		200	91.7	98.0	97.9	99.7
	9	50	84.1	96.0	95.1	98.6
		100	87.0	98.1	97.3	99.4
		200	92.2	99.4	98.7	99.9
5	3	50	82.0	93.1	94.4	95.1
		100	79.9	92.9	95.6	96.1
		200	80.6	87.0	95.5	96.6
	5	50	67.1	94.3	94.5	97.1
		100	66.1	93.7	96.0	97.6
		200	67.0	88.7	96.6	97.9
	7	50	57.5	92.2	94.2	98.0
		100	57.6	93.0	95.8	98.3
		200	62.2	91.6	97.4	99.3
	9	50	53.0	94.2	95.8	99.0
		100	55.2	94.0	96.2	99.1
		200	61.5	95.9	99.1	99.9

려하지 않는 경우에 비하여 relative bias의 절대값이 더 큰 값으로 얻어졌다. Fleiss의 방법은 여전히 다른 방법들에 비해 relative bias가 음수로 큰 반면, Conger의 방법은 평가자수와 평가대상자수가 작은 경우 relative bias가 양수로 크게 나타났다. 네 방법 모두 평가대상자수가 커질수록 relative bias가 0에 가깝게 수렴하고 있고 특히 Randolph와 Gwet의 방법이 더욱 그 현상이 뚜렷하다. Fleiss 방법은 이 경우에도 coverage rate이 다른 방법들에 비해 매우 낮고, 나머지 세 방법들은 비슷한 값을 갖는데, Randolph의 방법이 비교적 95%에 가장 가까운 값을 갖는 것으로 나타났다.

5. 수신증 자료

5.1. 자료 설명

수신증(hydronephrosis)이란 정상적인 요의 흐름이 폐색됨으로서 생기는 신장의 팽창 증상이다. 본 자료는 90명의 아동 환자의 좌우 신장을 4명의 방사선 전문의가 수신증의 진행 정도를 두 가지 평가 체계 즉, Society for Fetal Urology grading system(SFU)과 Onen's grading system(Onen)으로 각 두 번씩 평가한 결과이다. 두 체계는 모두 0에서 4점까지 다섯 범주로 평가할 수 있고 첫 평가를 한 후 적어도 24시간이 지난 후에 두 번째 평가를 하였다. 5.2절에서는 자료의 주변확률과 그에 따른 각 방법의 P_e, 일반화 카파통계량 κ값을 비교하였다. 실제 분석에서는, 두 번째 측정한 결과가 첫 번째 측정한 결과와 거의 비슷하였기 때문에 첫 번째로 평가한 결과의 자료만을 사용하였다.

Table 5.1. Marginal probability distributions and observed proportions of agreement (P_a) for the hydronephrosis data

자료	p_0	p_1	p_2	p_3	p_4	P_a
SFU_L	0.0472	0.0111	0.3194	0.2778	0.2444	0.7630
SFU_R	0.4194	0.1917	0.1917	0.1194	0.0778	0.7185
Onen_L	0.0556	0.3444	0.2278	0.1667	0.2056	0.7482
Onen_R	0.4611	0.3167	0.0972	0.0722	0.0528	0.8407

Table 5.2. Analysis results of the four generalized kappa statistics for the hydronephrosis data

	SFU_L				SFU_R			
	Fleiss	Conger	Randolph	Gwet	Fleiss	Conger	Randolph	Gwet
P_e	0.2535	0.2499	0.2000	0.1866	0.2697	0.2670	0.2000	0.1826
κ	0.6824	0.6837	0.7037	0.7086	0.6146	0.6160	0.6481	0.6557
SE	0.0373	0.0370	0.0356	0.0353	0.0394	0.0391	0.0383	0.0386
	Onen_L				Onen_R			
	Fleiss	Conger	Randolph	Gwet	Fleiss	Conger	Randolph	Gwet
P_e	0.2436	0.2414	0.2000	0.1891	0.3308	0.3304	0.2000	0.1674
κ	0.6670	0.6680	0.6852	0.6894	0.7624	0.7622	0.8009	0.8087
SE	0.0371	0.0369	0.0362	0.0362	0.0371	0.0371	0.0348	0.0344

5.2. 일반화 카파통계량 분석 결과

Table 5.1에 각 범주 별 주변확률의 결과로 평가를 하는 모든 경우의 수 즉, 아동 환자 90명을 방사선 전문의 4명이 모두 평가를 하므로 360인 경우의 수에서 각 범주에 아동 환자를 분류한 전문의의 비율을 제시하였다.

여기서 SFU와 Onen은 전문의가 평가하는 각 분류 체계의 유형이고, L과 R은 각각 좌 신장, 우 신장을 나타낸다. Table 5.1을 보면 0점부터 4점까지의 범주 간 주변확률 값의 차이가 존재하므로 균형적 주변분포가 성립하지 않는다고 할 수 있다.

따라서 이 자료에서는 각 방법에 따라 P_e와 κ값이 차이가 날 것을 예상할 수 있다. 관찰된 일치비율 P_a는 Onen 체계의 우 신장 자료가 0.8407로 가장 높았고, SFU 체계의 우 신장 자료가 0.7185로 가장 낮았다.

Table 5.2는 SFU와 Onen 분류 체계로 수신증 아동 환자 좌우 신장의 상태를 평가한 네 명의 방사선 전문의의 일치도를 네 가지 일반화 카파통계량 방법으로 분석한 결과이다. 각 방법 별 P_e, 일반화 카파통계량 κ와 표준오차(standard error; SE)를 제시하였다. 각 범주의 주변확률분포 결과를 보며 예상했듯이 각 방법 별 통계량 값이 차이가 존재함을 확인할 수 있었다. SFU와 Onen으로 평가한 체계 모두 Fleiss의 방법이 P_e가 가장 컸고, Gwet의 방법이 가장 작았다. Randolph의 방법은 범주수가 5이므로 모든 경우에서 P_e는 0.2로 동일하다. 따라서 일반화 카파통계량 κ는 Gwet의 방법이 가장 큰 값, 즉 P_a에 가장 가까운 값을 나타냈고, Fleiss 방법의 κ값이 가장 작았다. 표준오차는 거의 차이가 없었다.

6. 고찰 및 결론

본 논문에서는 일반화 카파통계량과 관련하여 주변확률 변화와 평가자수, 평가대상자수, 범주수에 따른 방법 비교에 대해 연구하였다. 실제 일치도 분석에서 카파통계량은 주변확률 변화에 따라 값이 의존적인 큰 문제점 때문에 이에 관한 많은 연구가 진행되고 있고, 보완하기 위한 몇몇 방법들이 제안되었다

(Brennan과 Prediger, 1981; Park과 Park, 2007; Gwet, 2010). 본 연구에서는 단순 카파통계량에서 제시된 두 가지 역설이 일반화 카파통계량에서는 어떻게 작용하는지 알아보고, 기존에 제시되어 있는 방법들을 다양한 조건에서 비교하여 주변분포에 덜 민감하고 모집단값을 가장 정확하게 추정하는 방법을 찾아보고자 하였다. 주변확률을 변화시키면서 균형적 주변분포와 주변동질성 여부에 따라 κ값의 양상을 방법 별로 살펴보았고, 평가대상자수, 평가자수, 범주수를 변화시키면서 방법들의 relative bias와 coverage rate을 비교하였다.

본 연구에서 비교한 네 가지 방법은 관찰된 일치비율 P_a를 구하는 공식은 모두 동일하지만 우연에 의해 기대되는 일치비율 P_e가 다르게 정의되어 서로 다른 일반화 카파통계량이 된다. Fleiss의 방법은 일반화 카파통계량 중 처음 제안되었고, 지금까지도 대표적으로 가장 많이 쓰이고 있는 방법이지만 주변분포에 κ값이 매우 의존적인 문제점이 있다. Conger는 Fleiss의 방법을 보정하여 좀 더 Cohen의 이차원적인 κ를 정확하게 확장시킨 통계량을 제시하였지만 계산이 상대적으로 복잡하고 평가자가 넷 이상인 경우에는 Fleiss의 방법과 거의 같아지며 역시 주변확률에 민감하게 작용한다. Randolph의 방법은 P_e를 범주의 역수로 정의하여 위의 두 방법과는 달리 주변확률에 영향을 받지 않는 큰 장점이 있고 범주수와 P_a가 같다면 평가자수와 상관없이 값은 동일하다. 마지막으로 Gwet의 방법은 통계량을 평가자 내 변동을 고려하여 P_e를 구한다. Randolph의 방법과 마찬가지로 κ값이 P_e에 민감하지 않도록 보완된 통계량으로 오히려 불균형적 주변분포일수록 높은 κ값을 보였다.

주변확률을 변화시켜 가면서 모의실험을 한 결과, 평가자수가 3 이상인 일반화 카파통계량은 평가자수가 2일 때 카파통계량을 구하는 경우 큰 문제점이 되었던 균형적 주변분포와 주변동질성 여부 중에서 균형적 주변분포 여부에만 강하게 영향 받는 것을 알 수 있었다. 예를 들어 평가자수가 3, 범주수가 2인 경우에 주변확률 분포가 (0.5, 0.5)인 경우는 균형적 주변분포, (0.1, 0.9)인 경우는 강한 불균형적 주변분포를 따른다고 할 수 있고 이 경우 P_a가 크더라도 Fleiss와 Conger의 방법의 일반화 카파통계량은 음수가 나오는 치명적인 문제점이 발생하였다. 하지만 이 문제는 평가자의 각 쌍이 서로 주변동질성을 만족하는지의 여부에 영향을 받는 것이 아니고 오로지 분포의 균형성 여부에만 영향을 받음을 확인하였다. 따라서 일반화 카파통계량에서는 주변동질성 여부는 중요하지 않고, Randolph와 Gwet의 방법이 다른 두 방법과는 달리 균형적 주변분포 여부에 민감하지 않고 특히 Gwet의 방법은 오히려 불균형 주변분포일 때 P_a에 가까운 κ을 도출할 수 있었다. 하지만 이 양상은 범주수가 커질수록 그 차이가 줄어들었다.

평가대상자수, 평가자수, 범주수를 변화시켜 가면서 모의실험을 한 결과 모든 조건에서 대체로 Randolph와 Gwet의 방법이 relative bias가 다른 두 방법에 비해 현저하게 작았고 특히 Randolph의 방법이 Gwet의 방법보다 relative bias가 다소 더 작았다. Fleiss와 Conger의 방법은 relative bias가 컸으며, 특히 Fleiss의 방법은 표본수가 작은 경우 다른 방법들보다 심하게 과소추정 되는 경향을 보였다. Conger의 방법도 평가자의 변동성을 고려한 모집단에서 추출한 경우 평가대상자수와 평가자수가 작을 때 심하게 과대추정이 되는 경향을 보이기도 했다.

Coverage rate은 Fleiss를 제외한 세 방법에서는 높게 추정되었으며 Randolph 방법이 실제 신뢰수준과 가장 비슷한 값을 도출하였다. Conger의 방법이 coverage rate이 높더라도 앞서 발견한 과소추정, 과대추정의 문제점 때문에 모집단값을 정확하게 추정하지 못한다고 여겨진다. Randolph와 Gwet의 방법은 주변확률에도 민감하게 반응하지 않으며 모집단값도 정확하게 추정함을 볼 수 있다.

본 논문에서 비교한 일반화 카파통계량의 네 가지 방법은 평가 결과의 범주를 명목형으로 가정하였다. 반면, 평가자가 두 명인 자료에서 평가의 결과가 순서형일 경우 가중 카파통계량을 이용하여 구할 수 있듯이 일반화 카파통계량도 순서형 자료인 경우를 고려할 수 있다 (Berry와 Mielke, 1988; Janson과 Olsson, 2001, 2004; Gwet, 2010). 이와 같은 순서형 자료에 적용 가능한 방법들을 비교하는 것도 흥미로운 연구가 될 것으로 보인다.

References

Berry, K. J. and Mielke, P. W. (1988). A generalization of Cohen's kappa, *Educational and Psychological Measurement*, **48**, 921–933.

Brennan, R. L. and Prediger, D. J. (1981). Coefficient kappa: Some uses, misuses, and alternatives, *Educational and Psychological Measurement*, **41**, 687–699.

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

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement of partial credit, *Psychological Bulletin*, **70**, 213–220.

Conger, A. J. (1980). Integration and generalization of kappas for multiple raters, *Psychological Bulletin*, **88**, 322–328.

Feinstein, A. R. and Cicchetti, D. V. (1990). High agreement but low kappa: 1. The problems of two paradoxes, *Journal of Clinical Epidemiology*, **43**, 543–549.

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

Gwet, K. L. (2008a). Computing inter-rater reliability and its variance in the presence of high agreement, *British Journal of Mathematical and Statistical Psychology*, **61**, 29–48.

Gwet, K. L. (2008b). Variance estimation of nominal-scale interrater reliability with random selection of raters, *Psychometrika*, **73**, 407–430.

Gwet, K. L. (2010). *Handbook of Inter-Rater Reliability*, 2nd edn. Advanced Analytics, LLC.

Janson, H. and Olsson, U. (2001). A measure of agreement for interval or nominal multivariate observations, *Educational and Psychological Measurement*, **61**, 277–289.

Janson, H. and Olsson, U. (2004). A measure of agreement for interval or nominal multivariate observations by different sets of judges, *Educational and Psychological Measurement*, **64**, 62–70.

Park, M. H. and Park, Y. G. (2007). A new measure of agreement to resolve the two paradoxes of Cohen's kappa, *The Korean Journal of Applied Statistics*, **20**, 117–132.

Quenouille, M. H. (1949). Approximate test of correlation in time-series, *Journal of the Royal Statistical Society, Series B, (Methodological)*, **11**, 68–84.

Randolph, J. J. (2005). Free-marginal multirater kappa: An alternative to Fleiss' fixed-marginal multirater kappa, *Paper presented at the Joensuu University Learning and Instruction Symposium*.

Scott, W. (1955). Reliability of content analysis: The case of nominal scale coding, *Public Opinion Quarterly*, **19**, 321–325.