Laboratório de Psicologia, 5(1): 81-90 (2007)

Acordo inter-juízes: O caso do coeficiente kappa

Ricardo Fonseca
Pedro Silva
Rita Silva
Instituto Superior de Psicologia Aplicada, Portugal

Resumo

Sempre que é preciso classificar um conjunto de dados num dado número de categorias, vários tipos de enviesamentos podem ocorrer. Com vista à sua minimização é frequente o recurso a mais do que um juiz para categorizar os mesmos dados, analisando-se posteriormente o seu grau de acordo e consequentemente a fiabilidade da classificação. Entre os vários índices de acordo inter-juízes mencionados na literatura, o coeficiente kappa (Cohen, 1960) é referido como o mais frequentemente utilizado quando as variáveis em estudo são nominais.

Neste artigo, procuramos descrever este coeficiente, apresentando a sua definição, pressupostos, fórmula, e ilustrando os passos para o seu cálculo. Exploramos também o seu desenvolvimento para kappa ponderado (Cohen, 1968). Por fim, algumas críticas feitas a este coeficiente de acordo inter-juízes são sumariamente discutidas.

Palavras-chave: Acordo inter-juízes, Coeficiente kappa, Kappa ponderado.

Abstract

Whenever one needs to classify a set of data in a given number of categories, several types of biases can occur. In order to minimize them, it's frequent to recourse to more than one judge to categorize the same data, analyzing afterwards the degree of their agreement and consequently the reliability of the classification. Among the several interrater agreement indexes mentioned in the literature, kappa coefficient (Cohen, 1960) is referred as the most frequently used when variables in study are nominal.

A correspondência relativa a este artigo deverá ser enviada para: Ricardo Fonseca, Instituto Superior de Psicologia Aplicada, Rua Jardim do Tabaco, 34, 1149-041 Lisboa, Portugal; E-mail: rfonseca@ispa.pt / psilva@ispa.pt / rita_ricot@yahoo.com

In this article, we attempt to describe this coefficient, presenting its definition, assumptions, formula, and illustrating the steps to its calculi. We also explore its development to weighted kappa (Cohen, 1968). Finally, some critiques made to this interrater agreement coefficient are briefly discussed.

Key words: Interrater agreement, Kappa coefficient, Weighted kappa.

Em Psicologia trabalhamos frequentemente com variáveis de tipo nominal, sendo a classificação dos nossos elementos de análise uma primeira etapa a considerar. A necessidade de classificar variáveis, enquadrando-as em *n* categorias, é algo comum em diversas situações, como por exemplo em contexto clínico (Tinsley & Weiss, 1975), análise de conteúdo (Kolbe & Burnett, 1991) ou em investigação na área da persuasão (Petty & Cacioppo, 1986).

O processo de classificação pode, em certos casos, ser bastante simples e pouco problemático, como quando optamos por classificar os participantes num estudo em categorias decorrentes da cor da sua camisola (e.g., vermelha, azul ou verde) ou da estação do ano em que fazem anos. No entanto, noutros casos onde a classificação já não é tão directa (e.g., categorizar respostas pelo número de ocorrências; categorizar tipos de pensamentos) a literatura na área dos métodos de investigação em Ciências Sociais (e.g., Hoyle, Harris, & Judd, 2002) alerta para a ocorrência de possíveis enviesamentos, nomeadamente por parte de quem realiza essa classificação. Aspectos como a sua motivação, personalidade, experiência de codificação ou outros factores externos presentes no momento de classificação podem contribuir para classificações menos correctas. Neste sentido, tal como quando utilizamos escalas de medida, também ao nível dos processos de classificação nominal o conceito de fiabilidade da mensuração deve ser assumido como uma das preocupações de base para quem conduz este tipo de investigações. Deste modo, avaliar e assegurar a consistência da medida referente ao processo de classificação é algo fundamental pois informa sobre a objectividade deste processo, a partir do qual se vão basear as conclusões e/ou análises subsequentes.

De modo a procurar controlar ou minimizar estes enviesamentos podem ser tomadas algumas medidas preventivas. Por exemplo, relativamente à codificação de entrevistas semi-estruturadas, pode optar-se por uma exaustiva fase de treino de codificação e uso de um manual detalhado de codificação[1]. Mas para além das estratégias preventivas, existem outras soluções pelas quais se pode optar. No momento posterior à classificação das unidades em análise nas diversas categorias, é frequente optar-se por uma estratégia que avalie a objectividade dessa classificação a partir de um grau específico de concordância entre dois ou mais elementos avaliadores (*juízes*) **– o acordo inter-juízes.**

O objectivo do presente artigo centra-se precisamente nessa procura de consistência nas avaliações e na obtenção de um índice de fiabilidade das classificações. De entre alguns coeficientes referidos na literatura (e.g., o coeficiente *S* de Bennet, Alpert e Goldstein, o Π de Scott ou o α de Aickin – ver Fleiss, 1981; Zwick, 1988 & Hsu & Field, 2003) desenvolveremos o coeficiente de Kappa, proposto por Cohen (1960), dando a conhecer (ou relembrando) ao leitor aquele que é o coeficiente mais utilizado quando é necessário classificar dados em categorias nominais (Chen & Krauss, 2004), tendo sido citado mais de 2000 vezes na literatura em psicologia (Hsu & Field, 2003). Assim, mais do que proceder a uma comparação exaustiva entre os tipos de índices existentes, descreveremos detalhadamente o kappa de Cohen – a sua definição, pressupostos, fórmulas e aplicação prática. Chamamos a atenção para o facto de que este artigo não pretende apresentar novos desenvolvimentos ou críticas a este índice de acordo, mas sim introduzir leitores ainda não familiarizados nem especializados neste procedimento estatístico.

[1] Para desenvolvimentos consultar Bartholomew, Henderson, e Marcia (2000).

Índices de acordo inter-juízes: Apresentando o Kappa de Cohen (1960)

A percentagem de casos em que dois juízes concordam relativamente à classificação de um conjunto de itens num dado número de categorias é talvez o método mais simples para se aceder à fiabilidade das variáveis categoriais. No entanto, este método não tem em conta a proporção desse acordo que é devida ao acaso (Cohen, 1960; Smith, 2000), tendo vários autores sugerido diferentes índices de concordância que contemplam no seu cálculo essa porção de acordo devida ao acaso.

Cohen, na década de 60, viria a desenvolver o índice de κ como um coeficiente de concordância dos julgamentos de dois juízes para dados nominais, em alternativa aos coeficientes de fidelidade existentes para os dados em escalas de magnitude. Actualmente, o coeficiente de Kappa continua a ser amplamente utilizado, sendo os pressupostos básicos apresentados pelo autor para o seu cálculo: "(1) as unidades em análise são independentes; (2) as categorias da escala nominal são independentes, mutuamente exclusivas e exaustivas[2]; e (3) os juízes actuam independentemente*"* (p. 38). Cada juiz pode distribuir as unidades de análise pelas diferentes categorias livremente, partindo-se do princípio que ambos os juízes são considerados igualmente aptos para a realização da tarefa.

Os dados relativos às avaliações dos juízes são apresentados em tabelas de contingência n x n (semelhantes à Tabela 2, ver secção *ilustração dos passos para calcular o coeficiente kappa*) em que as células na horizontal (linhas) representam, por exemplo, os julgamentos do juiz 1 e as células na vertical (colunas) os julgamentos do juiz 2. O coeficiente κ pode então ser definido como a proporção de acordo entre os juízes após ser retirada a proporção de acordo devido ao acaso, exprimindo-se na seguinte fórmula:

Para proporções:

$$\kappa = \frac{Po - Pa}{1 - Pa} \tag{1}$$

ou para Frequências:

$$\kappa = \frac{\Sigma \text{ fa} - \Sigma \text{ fe}}{\text{N} - \Sigma \text{ fe}} \tag{2}$$

na qual *Po* (ou Σ fa) é a proporção de acordo observado (ou o acordo observado, no caso de se utilizarem as frequências observadas), ou seja a proporção de unidades que os juízes classificaram nas mesmas categorias, e *Pa* (ou Σ fe) a proporção de acordo devido ao acaso (ou o acordo esperado, no caso das frequências), ou seja a proporção de unidades classificadas pelos juízes nas mesmas categorias por mera coincidência.

A proporção total de acordo observado (*Po*) é obtida pela soma dos valores das células que se encontram na diagonal traçada na tabela (do canto superior esquerdo para o inferior direito), ao passo que a proporção de acordo devido ao acaso (*Pa*) é obtida através da soma do produto entre *total linha* x *total coluna* e divisão desse valor pelo total de observações também para todas as células da diagonal

[2] Chama-se a atenção para o facto do 2º pressuposto ser paradoxal, uma vez que categorias mutuamente exclusivas implicam que as suas frequências de instanciação estejam negativamente correlacionadas, logo não poderão ser independentes. Embora esta questão possa gerar alguma controvérsia, estes são os pressupostos apresentados originalmente por Cohen (1960), pelo que decidimos não incluir a sua discussão neste artigo.

traçada[3]. Todos os valores das células fora da diagonal representam a proporção de desacordo que, embora possa fornecer informação útil relativamente às fontes de discordância (Bakeman, 2000), não é considerada independentemente no cálculo final (Cohen, 1960).

O limite máximo de κ é 1, representando o acordo perfeito entre os juízes. Por outro lado, quanto mais próximo de 0 estiver o valor de κ, mais este sugere que o grau de acordo entre os juízes se deve ao acaso (i.e., quando $Po=P\alpha$, $\kappa=0$, i.e., o acordo é exclusivamente devido ao acaso). Apesar de não se verificar em grande parte das suas aplicações práticas, este índice pode também assumir valores negativos (até ao limite de -1), reflectindo graus de acordo inferiores aos esperados pelo acaso.

Em resposta à necessidade sentida por alguns investigadores em diferenciar o grau de desacordo entre as diferentes categorias (por exemplo, quando numa cotação de pensamentos se considera que o desacordo entre positivo e negativo é mais grave que o desacordo entre positivo e neutro ou entre negativo e neutro), o próprio Cohen (1968) desenvolveu o *kappa ponderado* (κw). A passagem de um coeficiente κ para um coeficiente κw permite atribuir diferentes pesos aos desacordos, tornando-se assim uma estatística de concordância preferível quando se quer classificar um conjunto de dados em categorias ordenadas (relembramos que o κ distingue apenas entre acordo e desacordo, destinando-se apenas a categorias nominais) (Schuster, 2004).

No cálculo de κw são utilizadas três matrizes n x n: uma para as frequências observadas (equivalente a *Po* da fórmula do kappa); outra para as frequências esperadas (equivalente a *Pa* da fórmula do kappa); e uma terceira matriz de dados para os pesos atribuídos aos desacordos.

A fórmula do κw é expressa da seguinte forma:

Para proporções:

$$\kappa \mathrm{w} = 1 - \frac{\Sigma\, vij\ poij}{\Sigma\, vij\ paij} \tag{3}$$

ou para Frequências:

$$\kappa \mathrm{w} = 1 - \frac{\Sigma\, vij\ faij}{\Sigma\, vij\ feij} \tag{4}$$

na qual *vij*=peso do desacordo atribuído a cada célula da matriz n x n, *poij*=proporção observada em cada célula (ou *faij*=frequência observada em cada célula) e *paij*=proporção esperada na célula devido ao acaso (ou feij=frequência esperada na célula devido ao acaso) (ver exemplo prático na secção "Ilustração dos passos para calcular o coeficiente kappa ponderado").

Num exemplo dado por Bakeman (2000) para a atribuição dos pesos aos desacordos, o valor atribuído às células da diagonal de acordo da matriz dos pesos é 0, indicando a concordância entre os juízes; para as células imediatamente ao lado da diagonal de acordo o valor atribuído é 1; as células ainda mais afastadas obtêm o valor 2, seguindo-se esta lógica sucessivamente (ver Tabela 1). Cohen (1968) sugere uma atribuição dos pesos que torna o κw equivalente à correlação intraclasses – todas as células da diagonal de acordo entre os juízes obtêm o valor 0; as células das duas diagonais imediatamente adjacentes obtêm o valor 1; as duas diagonais seguintes obtêm o valor 2^2=4; as duas

[3] Note-se que para uma matriz estar correctamente preenchida o total de observações das linhas e o total de observações das colunas deve indicar o total de observações.

diagonais adjacentes a estas últimas obtêm o valor 3^2=9, seguindo-se esta lógica pelo número total das categorias. Estas formas de atribuição dos pesos são apenas exemplos de como esta pode ser feita, pois como Cohen (1968) indica, "these (positive) weights can be assigned by means of any judgment procedure set up to yield a ratio scale" (p. 215).

Tabela 1

Exemplo de matriz de pesos dos desacordos sugerida por Bakeman (2000)

	A	**B**	**C**	**D**
A	0	1	2	3
B	1	0	1	2
C	2	1	0	1
D	3	2	1	0

A forma de se computar a variância para o kappa e o kappa ponderado foi descrita por Fleiss, Cohen e Everitt (1969), sendo assim possível determinar se os kappas diferem ou não significativamente de zero. Mas tal como as correlações que apresentam uma baixa variância podem ser estatisticamente significativas desde que tenham casos suficientes, também valores de kappa que sejam bastante baixos podem ser significativos, devendo a decisão sobre se um valor de kappa é satisfatório ou não basear-se na magnitude absoluta do coeficiente (Bakeman, 2000). Apesar de não existir um valor objectivo específico a partir do qual se deva considerar o valor do kappa como adequado, encontram-se na literatura algumas sugestões que orientam normalmente esta decisão, destacando-se a proposta de Fleiss (1981):

<.40	pobre
.40-.75	satisfatório a bom
>.75	excelente

Ilustração dos passos para calcular o coeficiente kappa

Para ilustrar a utilização prática do kappa de Cohen (κ) considere-se o seguinte exemplo. Imagine que foi pedido a dois psicólogos especializados em Psicologia do Desenvolvimento que assistissem a 30 gravações de típicas *situações estranhas* e avaliassem o comportamento de cada bebé em função dos tipos de vinculação.

Passo 1: Construção da matriz de dados

Sendo que a categoria a avaliar é o tipo de vinculação e esta se divide, geralmente, em 3 estilos principais construiremos uma tabela de contingência 3x3, a partir das frequências observadas.

Tabela 2

Matriz de observações

	Psicólogo 2			
Psicólogo 1	Seguras	Ambivalentes	Inseguras	Total
Seguras	8 (2,93)	2	1	11
Ambivalentes	0	6 (2,40)	2	8
Inseguras	0	1	10 (4,77)	11
Total	8	9	13	30

Na Tabela 2, os valores contidos nas células diagonais indicam o número de observações de concordância entre os avaliadores (e.g., psicólogo 1 e 2 concordam que no comportamento de 8 bebés observados, estes pertenciam à categoria de vinculação segura). Fora destas células estão contidas todas as observação de discordância e de que tipo são (e.g., inseguras – seguras; ambivalentes – inseguras). O somatório de observações das linhas e o somatório de observações das colunas indica o total de observações, ou seja, os 30 registos dos comportamentos dos bebés.

Passo 2: Cálculo do total de frequências observadas (acordo observado) e o número total de frequências esperadas (acordo devido ao acaso)

O total de frequências observadas (acordo observado) é obtido pela soma de todos os valores contidos nas células diagonais. O total de frequências esperadas (acordo devido ao acaso) obtém-se somando as multiplicações entre o total da linha pelo total da coluna correspondente a cada célula de acordo e dividindo-se esse produto pelo total de observações (valores entre parênteses na Tabela 2). Vejamos o exemplo:

Σ fa=8+6+10=24
Σ fe=([11x8]/30+[8x9]/30+[11x13]/30)=2,93+2,4+4,77=10,1

Passo 3: Cálculo de Kappa

$$\kappa = \frac{\Sigma \text{ fa} - \Sigma \text{ fe}}{\text{N} - \Sigma \text{ fe}} = \frac{24 - 10{,}1}{30 - 10{,}1} = \frac{13{,}9}{19{,}9} = 0{,}69$$

Passo 4: Classificação do nível de acordo

De acordo com os intervalos habitualmente considerados na literatura, este índice revela uma boa concordância entre os dois psicólogos.

Decidindo-se pela aceitação deste valor, poderá prosseguir com a tentativa de resolver os casos de desacordos entre os dois psicólogos ou introduzir um terceiro avaliador para os desempatar, ou caso tal não seja necessário, utilizar apenas as unidades de análise em que houve acordo nas análises posteriores a realizar com estes dados.

Ilustração dos passos para calcular o coeficiente kappa ponderado

Imagine agora que o mesmo investigador sente a necessidade de diferenciar o grau de desacordo entre as diferentes categorias.

Passo 1: Atribuição dos pesos aos desacordos e cálculo das frequências esperadas para todas as células

Considerando que na cotação dos dois juízes o desacordo entre as categorias Seguras e Inseguras é mais grave que o desacordo entre Seguras e Ambivalentes ou entre Inseguras e Ambivalentes, o investigador atribui os pesos a esses desacordos conforme está descrito na Tabela 3 (valores entre parênteses rectos). Nesta tabela surgem também as frequências esperadas (ou seja, o acordo devido ao acaso) para todas as células, pois estes valores são necessários para o cálculo do κw.

Tabela 3

Matriz de observações com pesos dos desacordos

		Psicólogo 2		
Psicólogo 1	Seguras	Ambivalentes	Inseguras	Total
Seguras	8 (2,93 [0])	2 (3,33) [1]	01 (4,77) [2]	11
Ambivalentes	0 (2,13) [1]	6 (2,40) [0]	02 (3,46) [1]	08
Inseguras	0 (2,93) [2]	1 (3,33) [1]	10 (4,77) [0]	11
Total	8 (0,00) [0]	9 (0,00) [0]	13 (0,00) [0]	30

Passo 2: Cálculo do somatório de vij faij e de vij feij

Para se calcular Σ *vij faij*, somam-se os produtos entre os pesos e as frequências observadas de cada célula da matriz de dados; por sua vez, no cálculo de Σ *vij feij* somam-se os produtos entre os pesos e as frequências esperadas de cada célula. Neste caso, os cálculos seriam:

Σ *vij faij*=0(8)+1(2)+2(1)+1(0)+0(6)+1(2)+2(0)+1(1)+0(10)=7
Σ *vij feij*=0(2,93)+1(3,33)+2(4,77)+1(2,13)+0(2,40)+1(3,46)+2(2,93)+1(3,33)+0(4,77)=27,65

Passo 3: Cálculo do Kappa ponderado

$$\kappa w = 1 - \frac{\Sigma \text{ vij faij}}{\Sigma \text{ vij feij}} = 1 - \frac{7}{27{,}65} = 1 - 0{,}25 = 0{,}75$$

Passo 4: Classificação do nível de acordo

Tendo mais uma vez em conta os intervalos que são indicados na literatura, valor 0,75 representa um excelente nível de concordância entre os dois juízes, decidindo-se mais uma vez pela aceitação do mesmo. Uma vez aceite o valor, deverão seguir-se as mesmas acções referidas no passo 4 do exemplo do cálculo de *kappa*.

Recursos disponíveis para computar o coeficiente kappa

Os exemplos apresentados acima são relativamente simples para calcular o coeficiente de kappa, uma vez que apresentam apenas três categorias pelas quais se distribuíram as 30 observações, sendo assim

possível e fácil calcular este coeficiente manualmente. Mas nem sempre é este o caso. São vários os estudos em que o número de observações se aproxima ou ultrapassa as centenas e em que as categorias excedem a quantidade escolhida para os nossos exemplos. Nestes casos, os investigadores podem recorrer a software específico para o cálculo do acordo inter-juízes ou a programas estatísticos que incluem estes procedimentos.

Relativamente aos programas estatísticos que incluem o cálculo do coeficiente de kappa, encontramos o software Simstat, disponível através da Provalis Research (http://www.provalisresearch.com), e o SPSS (Statistical Package for the Social Sciences), disponível em http://spss.com.

No web site desenvolvido por Lombard, Snyder-Duch e Bracken (2005) existe informação sobre todos os softwares para o cálculo do kappa de Cohen que referimos, incluindo informação sobre como os obter e sobre como organizar os dados para serem analisados com cada um dos programas referidos. Através da página da Internet http://www.cosmion.net/jeroen/software/kappa/ desenvolvida por Geertzen (2006) é ainda possível calcular o coeficiente kappa *online*, estando aqui disponível um programa para o cálculo desta estatística no caso em que existem mais de dois juízes.

Limitações, críticas e comentários finais

Apesar do κ de Cohen ser o índice de concordância entre juízes mais utilizado e referido na literatura quando as variáveis são nominais, a sua aplicação é restrita às situações em que dois juízes avaliam todas as unidades de análise num determinado número de categorias. No entanto, em diversas ocasiões surge a necessidade de recorrer a medidas de acordo inter-juízes que vão para além destas condições, encontrando-se na literatura várias alternativas. Por exemplo, Fleiss (1971) sugere um desenvolvimento do κ para o caso em que, existindo múltiplos juízes, diferentes sub-grupos avaliam diferentes unidades de análise, mantendo constante o número de codificações por unidade de análise. Posteriormente, outras propostas mais flexíveis têm sido avançadas para situações em que se recorre a mais do que dois juízes (e.g., Schouten, 1986; Berry & Mielke, 1988; Posner, Sampson, Caplan, Ward, & Cheney, 1990).

Adicionalmente às limitações referidas relativamente à sua amplitude de aplicação, o coeficiente *kappa* têm sido alvo de algumas críticas no que respeita aos seus pressupostos.

Um aspecto central, do qual derivam diversas das críticas apresentadas por vários autores (e.g., Brennan & Prediger, 1981; Zwick, 1988), reside na definição do acaso na fórmula do coeficiente κ, nomeadamente $Pa=\Sigma\ Pi\ Pj$ (em que i representa o total da coluna e j o total da linha, por exemplo). Outros índices de concordância que partilham com o kappa a sua fórmula genérica (Po-Pa/1-Pa), apresentaram definições diferentes para a proporção de acordo devido ao acaso (por exemplo, o Π de Scott). Esta escolha de $\Sigma\ Pi\ Pj$ para esse efeito pode ter um impacto substancial tanto na magnitude como na interpretação de um índice de acordo, apontando estes autores que o kappa de Cohen penaliza indevidamente os avaliadores que produzem valores marginais (valores das células fora da diagonal de acordo) idênticos, uma vez que o acordo devido ao acaso ($\Sigma\ Pi\ Pj$) aumenta com o acordo marginal. Contudo, outros autores (e.g., Hsu & Field, 2003) defendem esta definição, uma vez que destaca o kappa de outros coeficientes pois a sua aplicação não requer nem pressupõe homogeneidade ou uniformidade ao nível dos valores marginais.

Uma outra questão que se coloca é o facto da descrição comum do coeficiente kappa como medida de consenso corrigida para o acaso poder ser questionável, uma vez que a sua computação passa pela

estimativa da proporção de acordo devido ao acaso. No entanto, esta proporção só representa uma estimativa apropriada quando existe independência estatística dos juízes, o que implica à confirmação da hipótese nula num teste de hipóteses em que esta representa a possibilidade de que o consenso encontrado é igual ao consenso expectável pelo acaso. Deste modo, podem ser levantadas objecções relativamente à possibilidade de aceitação da hipótese nula (ver Wickens, 1989)[4], e como afirma Ubersax (1987), uma vez que *Pa* deriva da condição em que a hipótese nula é de acordo totalmente devido ao acaso, não é claro como a magnitude de kappa deve ser interpretada quando esta hipótese é rejeitada.

Não sendo o nosso objectivo descrever estas e outras críticas exaustivamente, pretendemos apenas introduzir e alertar o leitor para a discussão existente na literatura (uma sistematização de algumas destas críticas pode ser encontrada em Hsu & Field, 2003). Neste sentido, apesar das limitações mencionadas, o kappa de Cohen apresenta-se como uma estatística de grande utilidade na validação de uma classificação, recomendando-se a sua utilização.

Bibliografia

Bakeman, R. (2000). Behavioral observation and coding. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 138-159). New York: Cambridge University Press.

Bartholomew, K., Henderson, A., & Marcia, J. (2000). Coded semistructured interviews in social psychological research. In H. Reis & C. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 286-312). New York: Cambridge University Press.

Berry, K., & Mielke, P. (1988). A generalization of Cohen's kappa agreement measure to interval measurement and multiple raters. *Educational and Psychological Measurement, 48*, 921-933.

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

Chen P., & Krauss, A. (2004). Interrater agreement. In M. Lewis-Beck, A. Bryman, & T. Liao (Eds.), *The sage encyclopedia of social science research methods* (vol. 2, pp. 511-513). Thousand Oaks, California: Sage Publications, Inc.

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 with provision for scaled disagreement or partial credit. *Psychological Bulletin, 70,* 213-220.

Fleiss, J. (1981). *Statistical methods for rates and proportions* (2th Ed.). New York: John Wiley & Sons.

Fleiss, J., Cohen, J., & Everitt, B. (1969). Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin, 72,* 323-327.

Geertzen, J. (2006). Cohen's kappa for more than two annotators with multiple classes. Consultado em 14 de Maio, 2007, através de: http://www.cosmion.net/jeroen/software/kappa/.

Hoyle, R. Harris, M., & Judd, C. (2002). *Research method in social relations* (7th Ed.). Pacific Grove, CA: Wadsworth Publishing.

[4] Agradecemos aos revisores deste artigo a chamada de atenção para esta questão.

Hsu, L., & Field, R. (2003). Interrater agreement measures: Comments on Kappan, Cohen's Kappa, Scott's Π, and Aickin's α. *Understandig Statistics, 2,* 205-219.

Kolbe, R., & Burnett, M. (1991). Content-analysis research: An examination of applications with directives for improving research reliability and objectivity. *Journal of Consumer Research, 18,* 243-250.

Lombard, M., Snyder-Duch, J., & Bracken, C. (2005). Practical resources for assessing and reporting intercoder reliability in content analysis research projects. Consultado em 14 Maio, 2007, através de: http://www.temple.edu/mmc/reliability/#How%20should%20researchers%20calculate%20intercoder%20reliability%20What%20software%20is%20available.

Petty, R., & Cacioppo, J. (1986). *Communication and persuasion: Central and peripheral routes to attitude change.* New York: Springer-Verlag.

Posner, K., Sampson, P., Caplan, R., Ward, R., & Cheney, F. (1990). Measuring interrater reliability among multiple raters: An example of methods for nominal data. *Statistics in Medicine*, *9*, 1103-1115.

Schouten, H. (1986). Nominal scale agreement among raters. *Psychometrika, 51*, 453-466.

Schuster, C. (2004). A note on the interpretation of weighted kappa and its relations to other rater agreement statistics for metric scales. *Educational and Psychological Measurement, 64,* 243-253.

Smith, C. (2000). Content analysis and narrative analyses. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 313-335). New York: Cambridge University Press.

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

Ubersax, J. (1987). Diversity of decision-making models and the measurement if interrater agreement. *Psychological Bulletin, 101,* 140-146.

Wickens, T. (1989). *Multiway contingency tables analyses for the social sciences.* Hillsdale, NJ: Erlbaum.

Zwick, R. (1988). Another look at interrater agreement. *Psychological Bulletin, 103,* 374-378.

Submissão: 28/02/2007 *Aceitação:* 06/06/2007