LA MEDICIÓN DE LA FIABILIDAD INTERJUEZ EN LA CODIFICACIÓN DE PREGUNTAS ABIERTAS: UNA PROPUESTA METODOLÓGICA

Francisco Muñoz Leiva
Francisco Montoro Ríos

RESUMEN

En el proceso de codificación de preguntas de respuesta abierta, un aspecto fundamental es la evaluación de la fiabilidad interjuez. En el presente trabajo se establece una propuesta metodológica para este proceso, así como un inventario de medidas de la concordancia o fiabilidad de la codificación. En la parte final, a partir de datos reales, se modeliza el comportamiento de tres coeficientes de fiabilidad entre codificadores, el coeficiente *Kappa* (*K*) de Cohen, el □ de Krippendorff y el *Ir* de Perreault y Leigh, en función del número de jueces implicados. Los resultados obtenidos muestran la importancia de esta variable en las valoraciones de la fiabilidad interjuez, así como la mayor adecuación del □ de Krippendorff y el *Ir* de Perreault y Leigh para las investigaciones de marketing.

PALABRAS CLAVES: Fiabilidad interjuez, Preguntas abiertas, Coeficientes de concordancia.

ABSTRACT

In the process of coding of open-ended questions, the evaluation of interjudge reliability is a critical issue. In this paper, we found a methodological proposal for this process, and a inventory of measurements of interjudge reliability. In the final part, using real data, the behaviour of three coefficients of reliability among coders, the Cohen's *K*, the Krippendorff's □ and the Perreault and Leigh's Ir coefficients are patterned, in terms of number of judges implicated. The outcome evidence the importance of this variable in the valuations of interjudge reliability, as well as the higher adequacy of Perreault and Leigh's Ir and Krippendorff's □ for the marketing research.

KEY WORDS: Interjudge reliability, Open-ended questions, Concordance coefficients.

1. INTRODUCCIÓN: EL PROCESO DE CODIFICACIÓN DE PREGUNTAS DE RESPUESTA ABIERTA EN LAS INVESTIGACIONES DE MARKETING

En la actualidad, el creciente interés por profundizar en el conocimiento de las motivaciones, conscientes e inconscientes, y en otros aspectos subyacentes que influyen en la conducta del individuo, ha generado una mayor atención hacia técnicas de investigación social cualitativas. No obstante, los estudios de mercado basados en técnicas cuantitativas (entrevistas personales, telefónicas, postales, 'mystery shopper' e Internet) suponían, en el año 2002, una parte muy importante del total de facturación del sector en España, con un 42% del total frente al 17% y al 41% correspondientes a estudios cualitativos y continuos respectivamente (Alòs, 2003).

Las entrevistas estructuradas utilizadas en investigaciones cuantitativas permiten la inclusión de preguntas abiertas que dejan autonomía de expresión al individuo, proporcionando por tanto un tipo de información de carácter eminentemente cualitativo. Las respuestas pueden ser recogidas por el entrevistador de acuerdo a un esquema de codificación previamente establecido (Fontana y Frey, 1994; p. 363).

En este tipo de preguntas de respuesta abierta, el individuo, al no estar sujeto a respuestas obligadas, puede expresar matizaciones y extenderse en explicaciones, ganando, de esta forma, en profundidad. Lo anterior amplía la diversidad de respuestas, sobre todo si tenemos en cuenta que no todos los encuestados tienen la misma capacidad de expresión ni un mismo estilo, lo que por otra parte constituye una potencial fuente de error. Si, por añadidura, las preguntas abiertas se formulan en una entrevista personal, resulta difícil registrar y sintetizar lo que se quiere decir (Luque, 1997; pp. 126-127; Lehmann *et al*., 1998: 178-179).

Las respuestas obtenidas mediante preguntas abiertas por lo general se traducen, tras el proceso de codificación, en una escala de tipo nominal, la cual permitirá identificar elementos diferentes, o denotará la pertenencia a una clase mediante una correspondencia unívoca, de forma que todos los miembros de una clase estarán asociados al mismo número. Al carecer de algunas de las propiedades de los números, como el orden o el origen, las posibilidades de análisis estadísticos quedan limitadas al análisis de frecuencias y determinados test no paramétricos, una vez que la información es condensada de forma cuantitativa.

Por tanto, una tarea crucial en el proceso de análisis de este tipo de preguntas es precisamente la codificación de la multitud de respuestas obtenidas. Dicha codificación consiste en esencia en asignar un identificador a cada categoría de datos, tarea descrita por varios autores entre los que destacan Lincoln y Guba (1985), Bardin (1986), Strauss y Corbin (1990), Miller y Crabtree (1994), Miles y Huberman (1994) y Glaser y Strauss (1999). De forma particular, y si se pretende que los resultados obtenidos tengan validez desde el punto de vista científico, la codificación debería ser realizada con la participación de codificadores (jueces) independientes.

Un aspecto al que no se le ha dedicado especial atención por parte de los investigadores de marketing es precisamente la evaluación de la calidad de los datos nominales derivados de juicios cualitativos. Son varios los autores que proponen que todo informe de investigación de marketing debe incluir de forma explícita la estimación de la fiabilidad del proceso de codificación (Light, 1971; Perreault y Leigh, 1989; Rust y Cooil, 1994). La fiabilidad entre codificadores está relacionada con sus discrepancias en la aplicación de criterios de categorización de contenidos. Las cuestiones principales en la elección de un índice de acuerdo son (Kang et al., 1993):

1.- La sensibilidad a errores de codificación sistemáticos.
2.- La corrección de los acuerdos debidos al azar.
3.- La capacidad de soportar a múltiples jueces.
4.- La escala de medida.

En el presente trabajo se hará especial hincapié en el establecimiento de metodología para la codificación y categorización de preguntas abiertas, y su posterior evaluación de fiabilidad mediante una serie de coeficientes. Para ello se tomará como referencia la metodología propia del análisis de contenido, la cual originariamente se creó para la explotación de la información generada por técnicas cualitativas procedentes de la psicología (entrevistas en profundidad y sesiones de grupo). Asimismo, ha encontrado aplicación en la investigación de marketing, especialmente en estudios sobre el carácter informativo de los anuncios publicitarios (Royo y Bigné, 1994; Abernethy y Franke, 1996) y en sus aspectos epistemológicos y metodológicos (Holsti, 1969; Berelson, 1952; Holbrook, 1977; Kassarjian, 1977; Weber, 1985; Bardin, 1986; López-Aranguren, 1989; Krippendorff, 1997; Bigné, 1999). Finalmente, se emplean datos de una investigación real para ilustrar gráficamente el efecto del número de jueces utilizados sobre los valores obtenidos en dichos coeficientes de concordancia.

2. EL PROCESO DE CODIFICACIÓN DE PREGUNTAS ABIERTAS

Las fases que componen el proceso de análisis de preguntas abiertas aparecen resumidas en la Figura 1. Dicho proceso debe ser considerado en el contexto de un proyecto amplio de investigación comercial, esto es, debe ser antecedido por la planificación preliminar y el diseño de la investigación. Por otra parte, algunas de las etapas incluidas son usuales en otras técnicas de análisis de información cuantitativa, o incluso cualitativa.

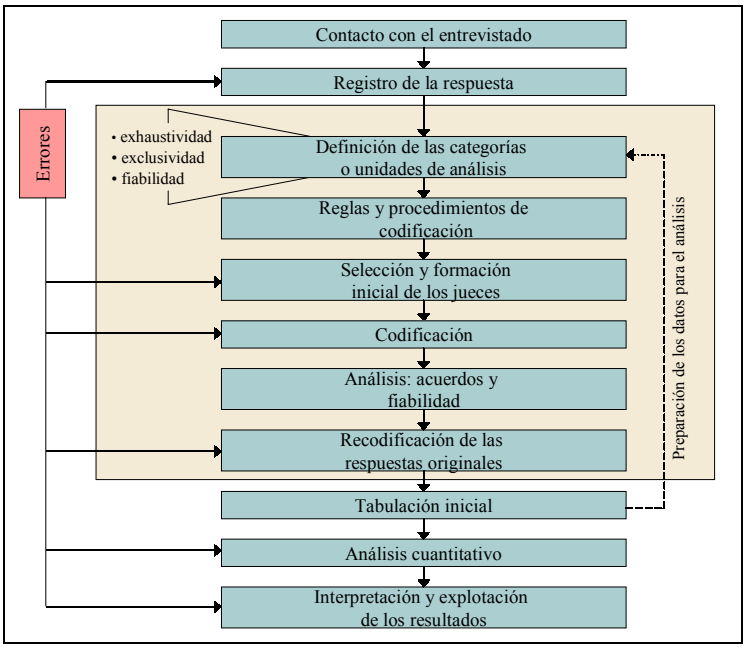

Figura 1: Proceso de tratamiento de preguntas abiertas (Elaboración propia)

El proceso comienza con el *registro de la respuesta* en el cuestionario, para, posteriormente, transcribirlo a una base de datos informatizada. En el registro de las respuestas es conveniente realizar una primera agrupación de las respuestas idénticas o muy similares, tanto por el entrevistador como por el encargado del tratamiento de la base de datos informatizada. En este sentido, el software funciona como un procedimiento descriptivo para obtener una visión de conjunto de la variedad, tipo o distribución de los datos, también aplicable en la tabulación preliminar previa al análisis (López Aranguren, 1989: 490). Tanto el registro de la respuesta como la trascripción de éstas a la base de datos pueden provocar errores no muestrales.

A continuación se *define una muestra de categorías* o unidades de análisis, que deben cumplir las condiciones de exhaustividad, exclusividad mutua y fiabilidad. Un análisis de frecuencias preliminar, unido a un examen de la literatura existente relacionada con el tema a tratar, puede facilitar bastante esta tarea. Tras esto se procede al *establecimiento de reglas y procedimientos de clasificación* estrictos que reduzcan la subjetividad de los jueces (Bigné, 1999).

En la siguiente fase se determina el número de jueces o codificadores independientes, y se realiza su selección y formación inicial, considerando que una inadecuada elección de los codificadores constituye otra fuente de sesgos o errores no maestrales, y que por tanto implica un menor nivel de fiabilidad del proceso.

En la *fase de codificación*, los jueces establecen una correspondencia entre las respuestas iniciales y las categorías presentadas en la segunda fase. Esta categorización se hace en base al significado coherente de cada

respuesta. La importancia de esta fase radica en la dependencia con la identificación inicial de categorías (Spiggle, 1994).

Posteriormente, al realizar el *análisis de los acuerdos y discrepancias* entre jueces, se establecen los criterios a utilizar en caso de empate (Rust y Cooil, 1994). Esta fase desemboca en la evaluación de la fiabilidad entre jueces, es decir, se cuantifican las coincidencias a la hora de asignar una determinada respuesta a la misma categoría mediante fórmulas basadas en el grado de acuerdo entre ellos (Bigné, 1999). Asimismo, el procedimiento resulta muy susceptible a la proliferación de errores no muestrales como consecuencia de fallos en la *recodificación de las respuestas* originales.

En la *tabulación preliminar*, resulta conveniente calcular la frecuencia de ocurrencia de las unidades de cada categoría. En el hipotético caso de haber obtenido demasiadas categorías con un escaso número de respuestas, puede resultar adecuado agruparlas en la categoría de "otras" respuestas, o en una nueva reagrupación que englobe a respuestas comunes[82].

Tras este examen previo se pasa a la fase de *análisis cuantitativo* de las categorías finales. Según los tipos de datos obtenidos se ofrece la posibilidad de aplicación de estadísticos descriptivos, o bien otros de tipo bivariante o incluso multivariante (Royo y Bigné, 1994; Krippendorff, 1997: 161-190).

Los aspectos que permiten la evaluación de cualquier proceso de codificación tienen que ver con: su objetividad (reglas y procedimientos de codificación, entrenamiento de los jueces, prueba de categorías y de sus definiciones, independencia de los jueces y número de jueces), su sistematización (si el proceso de codificación se utiliza para la prueba de hipótesis y/o teorías y si se describe el proceso de obtención de datos), el proceso de muestreo seguido y, finalmente, los índices de fiabilidad obtenidos (Kolbe y Burnnett, 1991). El presente artículo se centra en este último aspecto.

Se aprecia que se trata de un proceso laborioso y sujeto a errores a lo largo de sus diferentes fases. No obstante, si se aplican cuidadosamente los criterios y pasos correctos, se minimizan los posibles errores no muestrales que inciden y se garantiza el carácter científico del trabajo.

3. COEFICIENTES MÁS UTILIZADOS EN LA EVALUACIÓN DE LA FIABILIDAD ENTRE CODIFICADORES

Como se ha comentado anteriormente, el proceso de codificación de preguntas abiertas genera, normalmente, datos de carácter nominal. Los indicadores de fiabilidad más utilizados son la proporción de acuerdo entre jueces y otros coeficientes basados en esta concordancia, como es el caso de la □ de Cohen, el □de Krippendorff y el *CR* de Hosti (Kolbe y Burnnett, 1991). En el primer caso, el grado de acuerdo simple muestra una serie de inconvenientes: algunos acuerdos ocurren por azar, y para un menor número de categorías, es más probable que ocurra un acuerdo aleatorio, por tanto, la fiabilidad será mayor que la real (Rust y Coolí, 1994). Derivado de lo anterior, las medidas basadas en el grado de concordancia entre codificadores más utilizadas son:

1.- S de Bennett. Este primer coeficiente no puede ser utilizado en el caso de que operen más de dos jueces (Bennett y Goldstein, 1954):

[82] De otra forma puede ver dificultada la aplicación de medidas de asociación, como la Chi Cuadrado.

$$S = \left(\frac{F_o}{N} - \frac{1}{K}\right) \cdot \left(\frac{K}{K-1}\right) \quad (1)$$

donde,
F_o = número de acuerdos de los jueces.
N = número de elementos a codificar.
K = número de categorías total.

Se trata de un coeficiente aún más conservador que el *Kappa (K) de Cohen* mostrado a continuación. Sus valores oscilan entre 0 (ausencia de acuerdo) y 1 (acuerdo perfecto).

2.- Kappa (K) de Cohen. Este coeficiente (Cohen, 1960) se debe aplicar bajo los supuestos de independencia de codificadores y aleatoriedad de los efectos de estos (Hughes, y Garret, 1990).

$$\mathrm{K} = \frac{F_0 - F_c}{N - F_c} \quad (2)$$

donde,
F_o = número total de juicios coincidentes.
F_c = número de juicios coincidentes debidos al azar.
N = número total de juicios a emitir.

En el caso de concordancia perfecta, el coeficiente *Kappa* alcanza el valor 1, pero cuando las posibilidades de acuerdo coinciden con las obtenidas al azar, este coeficiente adquiere el valor 0. Los intervalos de la bondad del acuerdo aparecen reflejados en la siguiente tabla:

K de Cohen	Grado de acuerdo
Menor de 0	Sin acuerdo
0 – 0,2	Insignificante
0,2 – 0,4	Bajo
0,4 – 0,6	Moderado
0,6 – 0,8	Bueno
0,8 - 1	Muy bueno

Tabla 12: Intervalos de aceptación de la K de Cohen (Landis y Koch, 1977)

Este índice ha sido ampliamente criticado desde finales de los sesenta, debido a que fue pensado para aquellos juicios de psicología clínica en los cuales se supone que los jueces *a priori* asignarían pocos casos a enfermedades (categorías) "extrañas" (Perreault y Leigh, 1989; Hsu y Fied, 2003). Asimismo, se trata de un coeficiente bastante conservador debido a la forma de calcular los juicios coincidentes debidos al azar. Por estas razones, varios autores han introducido modificaciones en él a la hora de evaluar esta fiabilidad intercodificador (Fleiss, 1971; Krippendorff, 1971; Light, 1971; Herbert, 1977; Spitznagel y Helzer, 1985; Perrealult y Leigh, 1989; Hsu y Field, 2003). Las variaciones introducidas en el coeficiente se basan en la asunción de que las distribuciones marginales consideradas como "libres" son más apropiadas cuando no hay una razón previa para esperar una distribución marginal específica, como ocurre en el caso de los estudios de opinión (Perreault y Leigh, 1989).

3.- Coeficiente (CR) de Fiabilidad de Holsti. La formulación de este coeficiente fue planteada inicialmente por Holsti (1969: 140):

$$CR = \frac{2 \cdot M}{N_1 + N_2} \tag{3}$$

donde,

M = número de juicios en los que los evaluadores coinciden.

N_i = número de decisiones de codificación hechos por cada juez.

El coeficiente oscila entre 0 (total desacuerdo) y 1 (total acuerdo).

4.- Alpha (α) de Krippendorff. Se considera que se ha obtenido un nivel de concordancia aceptable si se obtienen valores por encima de 0,75, según la siguiente fórmula:

$$\alpha = 1 - \frac{D_0}{D_c} \tag{4}$$

D_o denota la proporción de desacuerdo observado y D_c la de desacuerdo esperado cuando la codificación de unidades es atribuible al azar. Posteriormente, Krippendorff (1980) generaliza el coeficiente a múltiples jueces y la existencia de datos perdidos con la siguiente expresión:

$$\alpha = 1 - \frac{D_0}{D_c} = \frac{(n-1) \cdot \sum_{c,u} \frac{n_{cc}}{m_u - 1} - \sum_c n_c (n_c - 1)}{n \cdot (n-1) - \sum_c n_c (n_c - 1)} \tag{5}$$

donde,

n = número total de decisiones o juicios emitidos por al menos dos jueces.

n_c = número de veces en que los jueces utilizaron la categoría c.

n_{cc} = número de juicios concordantes en la pareja c-c.

m_u = número de juicios en cada unidad de análisis u.

5.- Índice pi (π) de Scott. En base a la fórmula expuesta más abajo, se considera un nivel de acuerdo aceptable si el valor obtenido con este coeficiente es superior a 0,75. Cuando se utilizan dos jueces este índice es asintóticamente igual al coeficiente α *de Krippendorff* y al *K de Cohen*.

$$\pi = \frac{\%\text{ de acuerdo observado } - \%\text{ de acuerdo esperado}}{1 - \%\text{ de acuerdo esperado}} \tag{6}$$

Si hay más de dos codificadores en el estudio es posible calcular un coeficiente de fiabilidad compuesto, una vez calculado el π de Scott para cada par de codificadores (Holsti, 1969: 137).

$$\pi = \frac{\mathrm{N} \cdot (\text{acuerdo medio intercodificador})}{1 + [(\mathrm{N}+1) \cdot (\text{acuerdo medio intercodificador})]} \tag{7}$$

donde,

N = número de codificadores.

Este índice resulta útil cuando todas las unidades son codificados por todos los codificadores y resulta válido cuando el número de valores que toman las variables es pequeño, por ejemplo para tres categorías.

6.- I_r de Perreault y Leigh. El inconveniente de los anteriores coeficientes radica en que sus valores se ven influidos por el número de categorías de forma que a un menor número de categorías mayor es la probabilidad de obtener acuerdos al azar (y por tanto menor es el valor de aquellos). Con esta expresión el valor de la fiabilidad estimada aumenta conforme aumenta el número de categorías de respuestas, aunque de forma decreciente. El coeficiente de Perreault y Leigh (1989) consiste en una medida no ajustada a un contexto específico, lo que la hace adecuada para las investigaciones en marketing y estudios de opinión pública, dado que en estos casos no suele haber razón previa para esperar una distribución marginal específica, como sí plantearon los primeros investigadores que utilizaron el coeficiente *Kappa*. La formulación de este coeficiente es como sigue:

$$I_r = \left(\frac{F_o}{N} - \frac{1}{K}\right) \cdot \left(\frac{K}{K-1}\right)^{\frac{1}{2}} \quad si\ \frac{Fo}{N} \geq \frac{1}{K} \qquad (8)$$

$$I_r = 0 \qquad si\ \frac{Fo}{N} < \frac{1}{K}$$

donde,

F_o = número de juicios en los cuales los jueces coinciden.

N = número de juicios total.

K = número de categorías total.

Este indicador es posterior en el tiempo y, como se aprecia, representa la raíz cuadrada del *S* de Bennett. Con valores de fiabilidad I_r superiores de 0,9, el grado de acuerdo es alto, si oscila entre 0,7 y 0,9 se puede considerar como valor intermedio, mientras que es bajo (menor de 0,7) la asignación efectuada es conveniente solo para trabajos exploratorios. Alcanza el valor 1, cuando existe acuerdo interjuez perfecto, y el valor 0 cuando el número de acuerdos es menor o igual a los que serían esperados por azar si los jueces hubieran realizado, de forma totalmente aleatoria, las asignaciones a cada categoría (Rust y Cooil, 1994).

Kang *et al.* (1993) introducen una modificación en su formulación para generalizarlo a situaciones con más de dos jueces. Rust y Cooil (1994), a partir de los fundamentos del coeficiente α de Cronbach y la teoría de la Generalización de Heghes y Garret (1990), generalizan el índice de Perreault y Leigh al caso de utilización de múltiples jueces. Para ello, estiman la proporción de acuerdo interjuez, en la que cada uno selecciona correctamente, dada la proporción de acuerdo real (A) procedente de un testeo preliminar, de la siguiente forma:

$$\hat{p} = K^{-1}\left\{1 + \left[(K \cdot A - 1) \cdot (K-1)\right]^{1/2}\right\} \quad si\ A \geq \frac{1}{K} \qquad (9)$$

$$\hat{p} = \frac{1}{\mathrm{K}} \qquad si\ A < \frac{1}{K}$$

De esta forma, Rust y Cooli (1994) desarrollan una estimación de la fiabilidad (PRL) para aquellos casos con un número de categorías comprendido entre dos y cinco. Así, dado un número fijo de jueces y unos supuestos de partida, se estima la proporción de acuerdo que debería ser tenida en cuenta para garantizar una fiabilidad adecuada. Cuando existen solamente dos jueces, la medida PRL es equivalente a la medida de Perreault y Leigh.

4. TRABAJO EMPÍRICO

En el marco de un extenso estudio sobre comportamiento del consumidor[83] desarrollado en Julio de 2001 a nivel nacional, se recogieron respuestas relativas a los cuatro problemas sociales que a juicio del entrevistado más afectaban a la sociedad y a él personalmente. Esta formulación en el estudio al que hacemos referencia, como pregunta de introducción en un cuestionario más amplio.

Para un total de 703 cuestionarios correctamente obtenidos se obtuvieron un total de 3.601 respuestas diferentes que fueron agrupadas inicialmente en 508 problemas sociales. Posteriormente, y tras un análisis de frecuencias de las respuestas obtenidas, se describieron 29 categorías de respuesta.

Una vez redactado por escrito el contenido de cada una de las 29 categorías de respuestas identificadas inicialmente, se seleccionaron 6 jueces (independientes) que no habían participado en la planificación de la investigación, ni en su trabajo de campo. Tales jueces fueron debidamente formados mediante unas instrucciones redactadas por escrito, donde se recogían las categorías de respuesta y las principales cuestiones referidas al

[83] Este estudio ha sido realizado con el apoyo financiero prestado por el proyecto de investigación perteneciente al Plan Nacional de I+D financiado con fondos FEDER (refª 1FD97-0306).

procedimiento a seguir, junto con el formulario de codificación. Tras esto se procedió a la asignación de respuestas a categorías y su posterior evaluación de su fiabilidad.

En este estudio, nos centramos en tres indicadores de fiabilidad intercodificador: el coeficiente *K de Cohen y* el de Krippendorf por ser ambos de gran popularidad entre la comunidad científica, y el I_r de Perreault y Leight por tratarse de una alternativa a los anteriores y especialmente útil en investigaciones de marketing y estudios de opinión.

Una vez considerado el elevado número de jueces con el que se contó en el proceso de codificación y el número final de categorías unido al conservadurismo del primero, que limitan la aplicación de las medidas de la *K de Cohen*, resultó necesario introducir una serie de modificaciones en este coeficiente.

Para el cálculo de los juicios debidos al azar del coeficiente *Kappa*, se planteó su adaptación a nuestro caso particular a partir de la probabilidad de obtener una combinación con repetición CR de *m* elementos (en nuestro caso, 29 problemas), tomados de *n* en *n* (*n oscila entre* 2 y 6 jueces).

$$F_c = N \cdot \frac{1}{\frac{(m+n-1)!}{r!\cdot(m-1)!}} \tag{10}$$

Finalmente, el valor medio de los coeficientes (ver tabla siguiente) es obtenido a partir del promedio de todas las posibles combinaciones sin repetición (segunda columna) de *i* jueces tomados de n en n (desde 2 hasta 6):

$$\binom{6}{n} = \frac{6!}{n!\cdot(6-n)!} \tag{11}$$

Núm. de jueces	**Combinaciones de jueces**	**Coincidencias**	***Kappa revisada***		***de Krippendorff***	***Perreault y Leigh***
			Fc			
2	15	293	1,168	0,576	0,584	0,748
3	20	223	0,114	0,439	0,611	0,647
4	15	185	0,012	0,364	0,611	0,584
5	6	161	0,000	0,317	0,612	0,540
6	1	143	0,000	0,281	0,612	0,506

Tabla 13: Valor medio de los coeficientes de acuerdo de Cohen revisado, Krippendorff y Perreault y Leigh

El valor alcanzado por la *K de Cohen* para los seis codificadores fue de 0,281, dato no excesivamente bajo si se considera el número de jueces utilizados en la recodificación de la pregunta y la gran cantidad de categorías en las que había que agrupar las respuestas otorgadas a la pregunta abierta. No obstante, si se considera el promedio para todas las posibles combinaciones de pares de jueces, el coeficiente llega al 0,575, valor prácticamente incluido en el intervalo de bondad adecuada o moderada para juicios pareados; y al 0,439 para todas aquellas combinaciones de jueces tomados de tres en tres.

Vistos los valores correspondientes a los juicios debidos al azar (tabla 2) y la expresión de la Kappa de Cohen, se deduce claramente que conforme el número de jueces es mayor, el coeficiente tiende a la proporción de acuerdo simple.

La generalización obtenida por Krippendorff (1980) permite eliminar el efecto del número de jueces obteniéndose un valor medio de fiabilidad estable para la codificación.

El coeficiente de Perreault y Leigh, por su parte, obtiene mayores valores que el de Cohen en las diferentes combinaciones de jueces. De esta forma, la bondad del acuerdo puede considerarse como aceptable para el caso de 2 jueces, considerando siempre el elevado número de categorías consideradas en la codificación. Es a partir de este número de jueces, donde el Alpha de Krippendorff obtiene coeficientes mayores.

La representación gráfica nº 2 refleja el comportamiento de estos coeficientes en función del número de jueces incluidos.

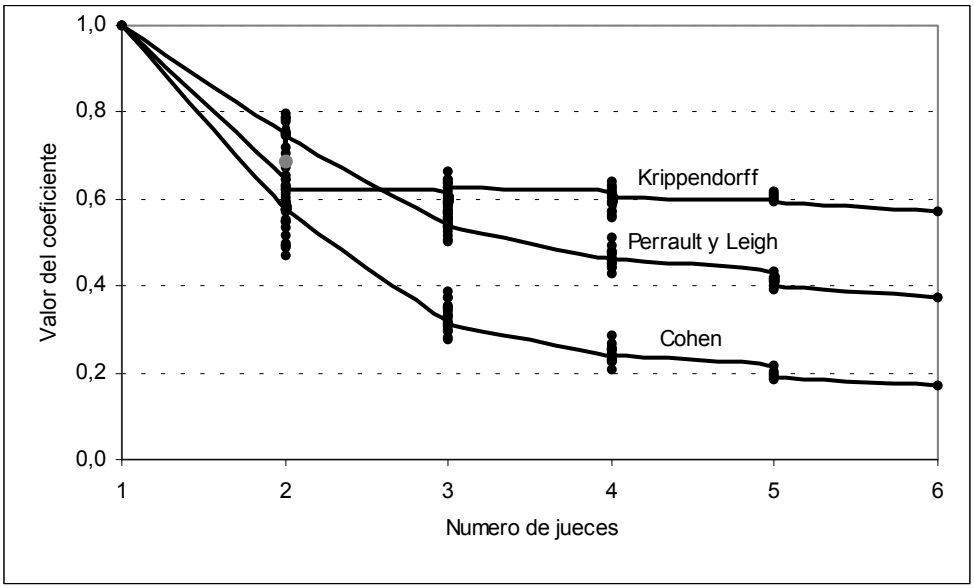

Figura 2: Valor de los coeficientes *K* de Cohen, de Scott, de Krippendorff I_r de Perreault y Leigh, en función del número de jueces

5. CONCLUSIONES

La fiabilidad entre jueces es percibida, a menudo, como la medida estándar de calidad del proceso de codificación. No obstante, valores altos de acuerdo entre los jueces podrían estar encubriendo parcas definiciones operacionales, así como, errores en el establecimiento de las categorías o en el entrenamiento de los jueces. En estas ocasiones, el coeficiente no tiene ningún significado (Kolbe y Burnnett, 1991). Se ha apreciado en todo el proceso la aplicación de criterios y procedimientos cuantitativos, descriptivos y sistemáticos, lo que acompañado de control exhaustivo y la revisión constante, permite garantizar el carácter científico de la técnica aplicada.

Una vez expuesto el proceso de codificación de preguntas de respuesta abierta y las mediciones más utilizadas en de su fiabilidad interjuez, se ha tratado de modelizar y generalizar el comportamiento de tres coeficientes, el *Kappa* (*K*) de Cohen, el de Krippendorff y el I_r de Perreault y Leigh, en función del número de jueces. Es necesario destacar que los resultados obtenidos están muy condicionados por el elevado número de categorías definidas en el estudio empírico.

A partir de los resultados, se demuestra la existencia de una relación inversa entre el valor alcanzado por los coeficientes y el número de jueces. Asimismo, y al igual que para los resultados obtenidos por Kang *et al.* (1993), el coeficiente de Perreault y Leigh obtiene valores mayores que el de la Kappa de Cohen. Esto, unido a la sistematización y rigurosidad seguida en todo el proceso, permite considerar este coeficiente con un mejor comportamiento (Rust y Coil, 1994) y, al mismo tiempo, más adecuado para aquellas situaciones en las que el número de categorías es elevado, como ocurre en un gran número de estudios de opinión con preguntas abiertas. Adicionalmente, el coeficiente de Krippendorff muestra una gran estabilidad en la medición de la concordancia otorgada por múltiples jueces. Por tanto, resulta útil para el caso de un elevado número de

codificadores, hecho que penaliza significativamente los dos coeficientes anteriores. Especialmente, para aquellas situaciones con más de dos jueces, vistos los inconvenientes del I_r de Perreault y Leigh.
Futuras investigaciones en este campo deberían recoger el comportamiento real de los coeficientes de fiabilidad en función de esta variable, para llegar a un cuadro teórico unificador aplicable en el proceso de codificación.

BIBLIOGRAFÍA

ABERNETHY, A. M. Y FRANKE, G. R. (1996): "The Information Content of Advertising: A Meta-Analysis", *Journal of Advertising*, vol. 25, nº 2, pp. 1-17.
ALÓS, J. S. (2003): "Industria de los Estudios de Mercado en España 2002", *Investigación y Marketing*, nº 80, pp. 76-78.
BARDIN, L. (1986). "*El Análisis de contenido*", Ed. Akal, S. A., Madrid.
BENNETT, E. M. Y GOLDSTEIN, A. C. (1954) : "Communications through limited response questioning", *Public Opinion Quaterly*, vol. 18, pp. 303-308.
BERELSON, B. (1952): "*Content Analysis in Communications Research*", Free Press, Glencoe (Ill.).
BIGNÉ, E. (1999): "El análisis de contenido", en SARABIA, F. J. (coord.): "*Metodología para la investigación en marketing y dirección de empresas*", Ed Pirámide, Madrid, pp. 203, 225.
COHEN, J. (1960): "A coeffcient of agreement for nominal scales", *Educational and Psychological Measurement*, vol. 20, nº invierno, pp. 37-46.
FLEISS, J. L. (1971): "Measuring nominal scale agreement among many raters", *Psychological Bulletin*, 76, 378-382.
FONTANA, A. Y FREY, J. H. (1994): "Interviewing. The Art of Science", en DENZIN, N. K. Y LICOLN, Y. S. (*eds.*): *Handbook of Qualitative Research*, Sage Publication, Thousand Oaks, CA.
GLASER, B. G. Y STRAUSS, A. L. (1999): "*The discovery of grounded theory: Strategies for qualitative research*", Aldine de Gruyter, New York.
HERBERT, L. (1977): "Kappa Revisited", *Psychological Bulletin*, vol. 84, nº 2, pp. 289-297.
HOLBROOK, M. B. (1977). "More on Content Analysis in Consumer Research", *Journal of Consumer Research*, vol. 4, nº 3, pp. 176-177.
HOLSTI, O. R. (1969): "*Content Analysis for the Social Sciences and Humanities*", Ed. Addison-Wesley, Reading, MA.
HSU Y FIED (2003): "Interrater Agreement Measures: Comments on Kappa, Cohen's Kappa, Scott's π and Aicking α", *Understanding statistics*, vol. 2, nº 3, p. 205.
HUGHES, M. A. Y GARRET, D. E. (1990): "Intercoder Reliability Estimation Approaches in Marketing: A Generability Theory Framework for Quantitative Data", *Journal of Marketing Research*, vol. 27, nº 2, pp. 185-95.
KANG, N.; KARA, A.; LASKEY, H.A.; SEATON, F.B. (1993): "A SAS MACRO for calculating interceder agreement in content analysis", *Journal of Advertising*, vol. 22, nº 2, pp. 17-28.
KASSARJIAN, H. H. (1977): "Content Analysis in Consumer Research", *Journal of Consumer Research*, vol. 4, nº 2, pp. 8-18.
KOLBE, R.H.; BURNETT, M.S. (1991): "Content-analysis research: An examination of applications with directives for improving research reliability and objectivity", *Journal of Consumer Research*, vol. 18, nº september, pp. 243-250.
KRIPPENDORFF, K. (1971): "Reliability of Recording Instructions: Multivariate Agreement for Nominal Data", *Behavioral Science*, vol. 16, nº 3, pp. 228-235.
KRIPPENDORFF, K. (1980): "*Content Analysis, an Introduction to Its Methodology*", Sage Publications, Thousand Oaks, CA.
KRIPPENDORFF, K. (1997): "*Metodología de análisis de contenido. Teoría y Práctica*", Paidós, Barcelona.
LANDIS, J.R., KOCH, G.G. (1977): "The Measurement of Observer Agreement for Categorical Data", *Biometrics*, vol. 33, pp. 159-174.
LEHMANN, R. L.; GUPTA, S. Y STECKEL, J. H. (1998): "*Marketing Research*", Addison-Wesley Educational Publishers Inc., Reading MA.
LIGHT, R. J. (1971): "Measures of response agreement for qualitative data: some generalizations and alternatives", *Psychological Bulleting*, vol. 76, nº 5, pp. 365-377.
LINCOLN, Y. S. Y GUBA, E. G. (1985): "*Naturalistic inquiri*", Sage Publications, Beverly Hills, CA.
LÓPEZ-ARANGUREN (1989): "El análisis de contenido", en GARCÍA, M.; IBÁÑEZ, J. Y ELVIRA, F. (Coord.): "*El análisis de la realidad social*: *Métodos y técnicas de investigación*", Alianza Editorial, Madrid.
LUQUE, T. (1997): "*Investigación de Marketing*", Ed. Ariel, Barcelona.
MILES, M. B. Y HUBERMAN, A. M. (1994): "*Quality Data Analysis. An expanded Sourcebook*". Sage Publications, Thousand Oaks, CA.
MILLER, W. L. Y CRABTREE, B. F. (1994): "Clinical Research", en DENZIN, N. K. Y LICOLN, Y. S. (*eds.*): "*Handbook of Qualitative Research*", Sage Publication, Thousand Oaks, CA.
PERREAULT, W. D. Y LEIGH E. L. (1989): "Reliability of Nominal Data Based on Quantitative Judgments", *Journal of Marketing Research,* 23 (May), 130-43.
ROYO, M. Y BIGNÉ, E. (1994): "Una aplicación del Análisis Multivariante al Contenido Informativo de la Publicidad en el Medio Televisión", *Investigación y Marketing*, vol. 45, nº julio, pp. 5-21.
RUST, R. T. Y COOIL, B. (1994): "Reliability measures for qualitative data: Theoryand implications", *Journal of Marketing Research*, vol. 31, pp. 1–14.
SPIGGLE, S. (1994): "Analysis and Interpretation of Qualitative Data in Consumer Research", *Journal of Marketing Research*. vol. 21 nº diciembre, p. 491-503.
SPITZNAGEL, E. L. Y HELZER, J. E. (1985): "A proposed solution to the base rate problem in the kappa statistic", *Archives of General Psychiatry*. vol. 42, nº. 7, pp. 725-28.
STRAUSS, A. Y CORBIN, J. (1990): "*Basic of qualitative research: Grounded theory procedures and techniques*". Sage Publications, Newbury Park, CA.

WEBER, R. P. (1985): "*Basic Content Analysis*", Ed. Sage, Newbury Park.