Quality & Quantity (2006) 40:519–537
DOI 10.1007/s11135-005-1093-6

Assessment of Interjudge Reliability in the Open-Ended Questions Coding Process

FRANCISCO MUÑOZ LEIVA*, FRANCISCO JAVIER MONTORO RÍOS and TEODORO LUQUE MARTÍNEZ
Department of Marketing and Market Research, Faculty of Economics and Business Sciences of the University of Granada, Campus Universitario La Cartuja s/n; 18071 Granada (Spain)

Abstract. In the process of coding open-ended questions, the evaluation of interjudge reliability is a critical issue. In this paper, using real data, the behavior of three coefficients of reliability among coders, Cohen's K, Krippendorff's α and Perreault and Leigh's I_r are patterned, in terms of the number of judges involved and the categories of answer defined. The outcome underlines the importance of both variables in the valuations of interjudge reliability, as well as the higher adequacy of Perreault and Leigh's I_r and Krippendorff's α for marketing and opinion research.

Key words: interjudge reliability; open-ended questions; concordance coefficients; coding process

1. Introduction

Academic circles and disciplines have traditionally been reticent about accepting qualitative research. Much criticism has been voiced concerning this type of studies: its use for exploratory ends, treating researchers as journalists more than scientists, or the presence of personal judgments (and prejudices) in the analysis process (Denzin and Lincoln, 1994; p. 4).

The current growing interest in investigating the motivations and other underlying aspects that influence an individual's conduct has meant that more attention is being paid to these techniques of social research. Thus, market studies based on qualitative techniques (group sessions or in-depth interviews) formed, in 2002, an important part of the sector's global turnover, totaling 15%, compared with 45% of quantitative-based studies (mainly through personal, phone, postal, 'mystery shopper' and *online* interviews) and 40% of panel studies (ESOMAR, 2003).

* Author for correspondence: F.M. Leiva, Department of Marketing and Market Research, Faculty of Economics and Business Sciences of the University of Granada, Campus Universitario La Cartuja s/n; 18071 Granada (Spain). Tel: +34-958-249603; Fax: +34-958-240695; E-mail: franml@ugr.es

However, the structured interviews used in quantitative research studies enable open-ended questions to be included, allowing the individual to express himself autonomously, thus providing a type of information that is eminently qualitative. The responses can be recorded by the interviewer using a previously-established coding process (Fontana and Frey, 1994; p. 363).

The very characteristics of open-ended questions mean that, on the one hand, they are mode difficult to code and analyze, but, on the other, there is a greater richness and depth en the responses. This is because, on not being limited to forced answers, the respondent is able to express nuances and provide more lengthy explanations. There is also a greater diversity of responses; above all if we bear in mind that not all the respondents have the same aptitude of expression or the same style, which, on the other hand, becomes a possible source of error. If, in addition to this, the open-ended questions are posed in a personal interview, it is even more difficult to record and synthesize what the respondent is trying to say (Luque, 1997; pp. 126–127; Lehmann et al., 1998: 178–179).

Responses obtained using open-ended questions are generally transferred, after the coding process, onto a nominal scale. This will help to identify different elements, or will indicate that an individual belongs to a certain class, by means of a univocal correspondence, such that all the members of one class will be associated to the same number. Since some of the properties of numbers, such as order or origin, are missing, the possibilities of statistical analysis are limited to calculating frequencies and percentages, as well as the carrying out of certain non-parametric tests. Taking all of this into account, and in spite of the limitations attributed to open-ended questions, the information obtained can be synthesized quantitatively, allowing it to be then treated statistically.

One aspect that marketing researchers have paid little attention to, is precisely the evaluation of the quality of the nominal data collected from qualitative judgments. Various authors propose that all marketing research reports should explicitly include the estimation of the coding process's reliability (Light, 1971; Perreault and Leigh, 1989; Rust and Cooil, 1994). In this sense, we should mention the results of a study presented by Hughes and Garrett (1990) on reliability analysis in marketing articles published from 1984 to 1987, which reveals that 46% of the articles developed qualitative judgments that were obtained from nominal scales.

Thus, in this study, we concentrate on the coding and classifying process of open-ended questions and, more specifically, on evaluating its reliability using the most-habitually used agreement coefficients. To do this, our reference point has been the methodology of content analysis, considered a research technique for the objective, systematic and quantitative description of the manifest of communication (Berelson, 1952: 18).

More recently, this tool has been used to exploit the information generated by the application of qualitative techniques that were borrowed from

Psychology, such as the in-depth interview and the group sessions. Likewise, it has also been applied in text analysis (Denzin and Lincoln, 1994; Miles and Huberman, 1994; Weitzman and Miles, 1995; Roberts, 2000), in the analysis of the informative aspects of publicity ads (Abernethy and Franke, 1996; Kassarjian, 1977; Lombard et al., 2002) and in the epistemological and methodological aspects of content analysis itself (Berelson, 1952; Holsti, 1969; Holbrook, 1977; Kassarjian, 1977; Krippendorff, 1980; Weber, 1985; Bardin, 1986; López-Aranguren, 1989; Krippendorff, 1997).

Finally, data from an actual research project is used to illustrate mathematically and generalize the effect on the values obtained in said agreement or concordance coefficients, in terms of both the number of judges used and of the categories described.

2. Evaluation of Interjudge Reliability in the Open-Ended Question Coding Process

A crucial task in the process of analyzing open-ended questions is, precisely, coding the multitude of responses obtained. This coding basically consists in attaching an identifier to each category of data. A process described by various authors, including Lincoln and Guba (1985), Bardin (1986), Strauss and Corbin (1990), Miller and Crabtree (1994), Miles and Huberman (1994) and Glaser and Strauss (1999). In particular, and if we want the results obtained to be scientifically valid, the coding should be carried out using independent coders (judges).

After preparing a sample of categories or units of analysis, in the coding phase the judges will establish a correspondence between the initial responses and these units. This classification is based on the coherent meaning of each response and on the assumption that the different judges are able to group each response, together in the same classifications (reliability). The importance of this phase lies in the dependence with the initial identification of categories (Spiggle, 1994).

On another note, in the task of recording the responses and placing them into groups, the use of a software package[1] (specialized or not) provides a descriptive procedure in order to obtain an overall vision of the variety, type or distribution of the data. This is also applicable in the tabulation carried out prior to the analysis (López-Aranguren, 1989: 490; Mckensen and Wille, 1999).

Analyzing the agreements and discrepancies gives us the interjudge reliability. That is, the quality of the research is quantified through formulae or numerical indices based on the level of agreement between them. An agreement occurs when the different judges coincide in placing a certain response in the same category. Therefore, the interjudge reliability is related to its discrepancies when applying content classification criteria.

Lombard et al. (2002) propose the patterns and models to be followed in order to calculate and present the intercoders reliability. Nonetheless, and as we shall see below, reliability analysis is a critical problem when multiple judges are used to assign codes (Kassarjian, 1977).

The paper of Hughes and Garrett (1990) reveals that only 13% of articles analyzed use acceptable measurements of the level of agreement among coders. The main questions in choosing an agreement index are (Kang et al., 1993): sensitivity to errors of systematic coding, correction of chance agreements, ability to support multiple judges and the measurement scale on which it can be applied.

The most frequently-used reliability indicators are interjudge agreement proportion and other measurements based on this concordance, such as Krippendorff's α and Holsti's *CR* (Hughes and Garrett, 1990; Kolbe and Burnnett, 1991; Kang et al., 1993; Riffe and Freitag, 1993). The simplest, most-easily calculated and understood indicator is the proportion of agreements between pairs of judges as regards the total number of judgments given. However, this measurement presents a group of disadvantages that they do it inappropriate to evaluate intercoder reliability (Hughes and Garret, 1990; Krippendorff, 1980). Concretely, some agreements occur by chance and, for a lower number of categories, a chance agreement is more likely, thus, the reliability will appear to be greater than it really is (Rust and Cooil, 1994). For this reason, other, more complex, concordance measurements have been developed. The most habitually used and their goodness intervals are listed in Table I.

Table II shows the main characteristics of the concordance coefficients mentioned in the text above.

It should be said that *Cohen's Kappa* (*K*) *Coefficient* must be applied using the assumptions that the coders are independent and that their effects are random (Hughes, and Garret, 1990). Ever since the end of the 1960s, this index has been widely criticized, since it was designed for clinical psychological judgments in which it is assumed that the judges, *a priori*, would assign very few cases to "strange" illnesses (categories) (Perreault and Leigh, 1989; Hsu and Fied, 2003). Thus, it is a useful coefficient when the set of response patterns are expected to be evaluated by comparison with an already-established standard.

The aforementioned reasons, along with its conservatism in calculating the random coinciding judgments, in its original formulation, have produced different variants of the *K* so as to adjust it to specific situations within a range of disciplines (psychology, sociology and marketing), when evaluating this intercoder reliability (Fleiss, 1971; Krippendorff, 1971; Light, 1971; Herbert, 1977; Spitznagel and Helzer, 1985; Perreault and Leigh, 1989; Hsu and Field, 2003). These modifications are based on

Table I. Concordance coefficients most habitually used

Coefficient	Expression	Agreement goodness intervals	
Bennett's *S* (1954)	$S = \left(\frac{F_o}{N} - \frac{1}{K}\right) \cdot \left(\frac{K}{K-1}\right)$		
	F_o = number of judge agreements.	Their values range from 0	
	N = number of elements to be coded.	(no reliability or agreement) to 1 (perfect	
	K = total number of categories.	reliability or agreement).	
		<0	No agreement[a]
Cohen's *K* (1960)	$K = \frac{F_o - F_c}{N - F_c}$	0–0.2	Insignificant
		0.2–0.4	Low
	F_o = total number of coinciding judgements.	0.4–0.6	Moderate
	F_c = number of coinciding judges due to chance.	0.6–0.8	Good
	N = total number of judgments to be given.	0.8–1	Very good
Holsti's *CR* (1969)	$CR = \frac{2 \cdot M}{N_1 + N_2}$		
	M = number of judgments in which evaluators coincide.	Ranges from 0 (complete disagreement)	
	N_i = number of coding decisions made by each judge.	to 1 (complete agreement).	
Krippendorff's (α) (1980)	$\alpha = 1 - \frac{D_o}{D_c} = \frac{(n-1) \cdot \sum_{c,u} \frac{n_{cc}}{m_u - 1} - \sum_c n_c (n_c - 1)}{n \cdot (n-1) - \sum_c n_c (n_c - 1)}$		
	D_o = proportion of disagreement observed.		
	D_c = proportion of disagreement expected when coding of units is put down to chance.		
	In second (generalized) expression and for nominal data:	In its original formulation, an	
	n = total number of decisions or judgments given by at least two judges.	acceptable level of concordance is considered	

Table I. Continued

Coefficient	Expression	Agreement goodness intervals	
	n_c = number of times judges used category c.		to have been obtained if values over 0.75 are
	n_{cc} = number of concordant judgments in pair c-c.		reached. Range of variation among the
	m_u = number of judgments in each unit of analysis u.		same values as the previous one.
Scott's *pi* (B) *(1955, in Hughes & Garret, 1990)*	$pi = \frac{PA_o - PA_e}{1 - PA_e} = \frac{N \cdot \overline{PA}}{1 + (N+1) \cdot \overline{PA}}$		
	PA_o = proportion of agreement observed.		
	PA_e = proportion of agreement expected.		
	In the composite reliability index, for the case of more than		A level of agreement is considered
	two coders:		acceptable is the value obtained
	N = number of coders.		is over 0.75.
	PA = Mean intercoder agreement.		
Perreault and Leigh's I_r (1989)	$I_r = \left(\frac{F_o}{N} - \frac{1}{K}\right) \cdot \left(\frac{K}{K-1}\right)^{\frac{1}{2}} \; si \; \frac{Fo}{N} \geq \frac{1}{K}$ $I_r = 0 \; si \; \frac{Fo}{N} < \frac{1}{K}$		
	F_o = number of judgments in which judges coincide.		
	N = total number of judgments.		
	K = total number of categories.	0.9 – 1	High reliability

		0.7 – 0.9	Intermediate reliability
		<0.7	Low reliability (exploratory work)
Perreault and Leigh's I_r modified, PRL (Rust & Cooil, 1994)	$\hat{p} = K^{-1}\left\{1 + [(K \cdot A - 1) \cdot (K - 1)]^{1/2}\right\} \quad si\, A \geq \frac{1}{K}$ $\hat{p} = \frac{1}{K} \quad si\, A < \frac{1}{K}$		
	A = proportion of agreements from one testing. K = total number of categories.	0.9 – 1	High reliability
		0.7 – 0.9	Intermediate reliability
		<0.7	Low reliability (exploratory work)

Source: Own production

[a] *Goodness intervals according to Landis and Koch (1977)*

Table II. Characteristics of concordance coefficients

Coefficient	Multiple judges	Chance agreements	Conservatism	Scale
Bennett's *S* (1954)	No	No	++	Nominal
Cohen's *K* (1960)	Yes	Yes	+	Nominal
Holsti's *CR* (1969)	No	No	=	Nominal
Krippendorff's (α) (1980)	Yes	Yes	−−	Nominal, ordinal, interval
Scott's *pi* (B)	Yes	Yes	+	Nominal
Perreault and Leigh's I_r (1989)	No	No	−	Nominal
Perreault and Leigh's I_r modified, PRL (Rust & Cooil, 1994)	Yes	No	−	Nominal

Source: Own production after Kang et al. (1993)

the assumption that the so-called "free" marginal distributions are more appropriate when there is no prior reason for expecting a specific marginal distribution, as in the case of opinion studies (Perreault and Leigh, 1989).

The Krippendorff, Scott and Cohen coefficients are based on the idea that, if the agreement level obtained is not significantly higher than that expected randomly, the researcher will conclude that the intercoder reliability has not been sufficiently adequate (Hughes and Garret, 1990; Krippendorff, 1997, p. 197).

When two judges are used, the Scott *pi* index is asymptotically equal to *Krippendorff's* α and to *Cohen's K*. If there are more than two coders in the study, it will be possible to calculate a compose reliability coefficient, once Scott's *pi* has been calculated for each pair of coders (Holsti, 1969: 137).

The disadvantage of the aforementioned coefficients lies in the fact that their values are influenced by the number of categories. Therefore, the less categories there are, the greater the probability of obtaining random agreements and, thus, the lower their value. In this sense, Perrault and Leigh's (1989) I_r grows when the number of categories grows, but in a decreasing rhythm. In conceptual terms, Perreault and Leigh's (1989) coefficient can be taken as the percentage of overall responses a judge could consistently code, taking into account the nature of the observations, the coding pattern and the category definitions, as well as the judge's inclinations and abilities. As we can see, it represents the square root of Bennett's *S*.

In terms of an adequate, generally approved, index, Perreault and Leigh's seems to be well adapted to many research conditions, though only for two evaluators (Grayson and Rust, 2001). Nonetheless, Rust and Cooil (1994), using the basics of the Cronbach α coefficient and Hughes and Garret's Generability Theory, extend the use of Perreault and Leigh's index to the case of multiple-judge participation, as long as there are less than five categories to be coded. Thus, given a fixed number of judges, the agreement proportion to be borne in mind, so as to guarantee an adequate reliability, is estimated.

To sum up this section, the aspects that allow any coding process to be evaluated are based on (Kolbe and Burnnett, 1991): its objectivity (coding rules and procedures, judge training, testing of categories and their definitions, independence of judges and number of judges), its systematization (whether the coding process is used to check hypotheses and/or theories and whether the data collection process is described), the sampling process followed and, finally, the reliability indices obtained.

It is, evidently, a process that is empirical, laborious and subject to non-sample errors throughout its different phases. However, if the correct criteria and steps are carefully applied, the possible non-sample errors made are minimized, thus ensuring the scientific nature of the work.

3. Empirical Application

In the context of a wider study on consumer behavior carried out on a national level in July 2001, responses were recorded concerning the four problems that, according to the respondent, most affected, on the one hand, society in general and, on the other, the respondent himself. This formulation, identical to that habitually used by the Spanish Center for Sociological Research, CIS (Spain) in its opinion polls, was used as an introductory question to a more extensive questionnaire in the study we are dealing with here.

For 703 correctly-received questionnaires, a total number of 3,601 different responses were obtained, which were initially divided up into 508 social problems. Subsequently, following a frequency analysis of the responses obtained, 29 response categories were described (see Appendix).

Once the content of each of the 29 initially-identified response categories was written up, 6 judges who had not taken part in the research design were selected. These judges were duly trained using written instructions, which included the response categories and main questions referring to the procedure to be followed, along with the coding form. These instructions helped them to work in an isolate manner, assigning each of the 508 problems to the 29 categories formulated.

Once the responses were assigned to the categories, their reliability was evaluated. In this study, we concentrated on four concordance indicators: the agreement proportion, *Cohen's K* coefficient (revised) and Krippendorf's α coefficient given the popularity of both among the scientific community, along with Perreault and Leigh's I_r, as an alternative to the others and of particular use in marketing research and opinion studies.

In transcribing the judges' evaluations, the preliminary tabulations and other data processing needed to calculate these coefficients, a spreadsheet was designed using Microsoft Excel. The final classification consisted of assigning the modal analysis unit to each original response, and the categories that showed a draw were selected randomly. Furthermore, calculating the coefficients for all the possible combinations of judges or coders was automatized thanks to a series of "macros' and forms programmed in Visual Basic language (see Table III). Finally, simulating the behavior of the coefficients was obtained in graph form using the computing tool STATISTICA 6.0 from the data provided by the spreadsheet.

In the agreement analysis phase, and bearing in mind the problem of the high number of categories and judges, together with the conservatism of Cohen's *K*, a series of modifications had to be introduced in order for it to be applied.

To calculate the *K* coefficient's chance judgments F_c, we decided to adapt it to our particular case based on the probability of obtaining a combination with a repetition (*CR*) of *m* elements (in our case, 29 problems), taken *n* by *n* (*n* ranges from 2 to 6 judges). For the set of N responses, F_c has the following expression:

$$F_c = N \cdot \frac{1}{CR(m,n)} = N \cdot \frac{1}{(m+n-1)!/n! \cdot (m-1)!} \tag{1}$$

Finally, the mean value of the coefficients (see table below) is obtained from the average of all the possible combinations without repetition (second column) of *i* judges (6) taken *n* by *n* (from 2 to 6):

$$C(6,n) = \binom{6}{n} = \frac{6!}{n! \cdot (6-n)!} \tag{2}$$

Cohen's K for the six coders was 0.281, a value that is not excessively low if we take into account the rigorousness and systematization followed in the coding process, the number of judges used in recoding the question and, the large amount of categories to which to assign. Nonetheless, if we consider the average of all the possible combinations of pairs of judges, the coefficient rises to 0.575, a value that can practically

Table III. Mean value of agreement proportion, revised Kappa coefficient, Krippendorff's coefficient and Perreault and Leigh's coefficient

No. of judges	Combinations of judges	Coincidences	Revised Kappa			Krippendorff's α	DesvEst*	Perreault & Leigh	DesvEst*
			F_c	K	DesvEst[a]				
2	15	293	1.168	0.576	0.051	0.584	0.052	0.748	0.036
3	20	223	0.114	0.439	0.040	0.611	0.036	0.647	0.032
4	15	185	0.012	0.364	0.028	0.611	0.025	0.584	0.025
5	6	161	0.000	0.317	0.017	0.612	0.015	0.540	0.016
6	1	143	0.000	0.281	0.000	0.612	0.000	0.506	0.000

[a] *Standard deviation (population).*

be included in the adequate goodness interval. This same coefficient reaches 0.439 for all those combinations of judges taken 3 by 3.

Looking at the values for chance judgments and Cohen's Kappa expression (Table II), it is clearly deduced that, the more judges (and categories) there are, the more the coefficient tends toward a simple agreement proportion.

The generalization obtained by Krippendorff (1980) means that the effect of the number of judges can be eliminated, thus obtaining a stable mean reliability for coding, in spite of certain dispersion in the different combinations of few coders (two or three). For the case of two judges, the values of Cohen's K and Krippendorff's α coefficients coincide.

On another note, the Perreault and Leigh coefficient obtains higher values than that of Cohen in the different combinations of judges. Thus, the influence of the number of categories on the intercoder reliability is eliminated and the goodness of the agreement can be considered as acceptable for the case of 2 judges, always taking into account the high number of categories used in the coding. Above 2 judges, Krippendorff's Alpha obtains greater coefficients.

Figure 1 reflects the behavior of these coefficients in terms of the number of judges included. Scott's *pi* coefficient shows a value of 0.69 for the case of measuring agreements between two judges.

From the above analyses it follows that there is an inverse relationship between the number of judges and the reliability measurement. In the same

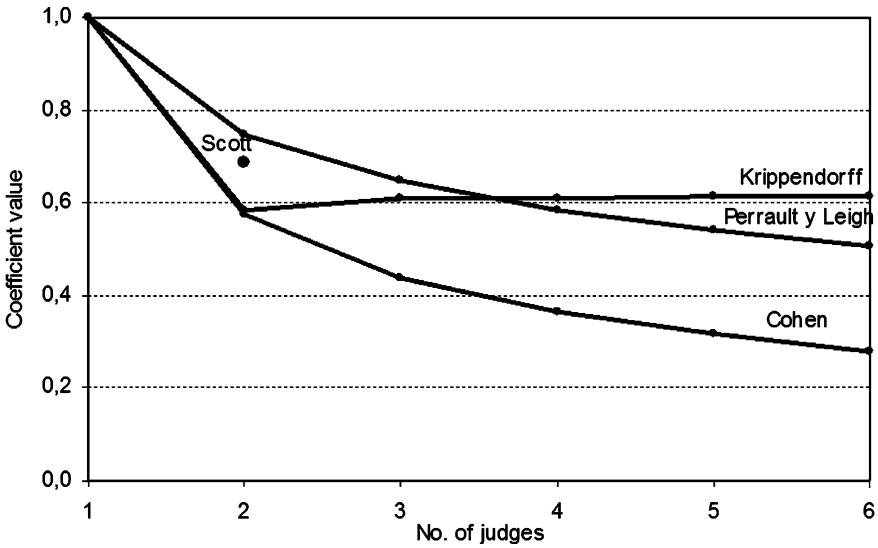

Figure 1. Value of Cohen's K coefficient, Scott's *pi (B)* coefficient, Krippendorff's α coefficient and Perreault and Leigh's coefficient, in terms of number of judges.

Table IV. First-order interpolation equations calculated.

No. of judges	Krippendorff's α		
	O_{cc}	$\sum_c n_c(n_c-1)$	K and Ir
2	$A_o = -38.294\,c + 2930.294$	$A_s = 11463.714\,c + 246092.286$	
3	$A_o = -31.912\,c + 2441.912$	$A_s = 5244.714\,c + 252311.286$	
4	$A_o = -25.529\,c + 1953.529$	$A_s = 135.838\,c + 257420.162$	
5	$A_o = -19.128\,c + 1463.728$	$A_s = -3757.496\,c + 261313.496$	
6	$A_o = -12.765\,c + 976.765$	$A_s = -6751.543\,c + 264307.543$	$A = -12.75\,c + 512.75$

way, Perreault and Leigh (1989) obtain a practically linear relationship in the relation between the number of categories and the intercoder reliability for four or more categories. In order to evaluate the relationship between the concordance indicators and the number, both of judges and categories, the values for this latter variable (less than 29 categories) were interpolated. The general linear interpolation equation (first order) for two initial (c_1) and final (c_2) coding units and their corresponding agreement values/coincidences (images) takes the following expression:

$$A(c) = A(c_1) + \frac{A(c_2) - A(c_1)}{c_2 - c_1} \cdot (C - c_1) \tag{3}$$

Finally, we obtained the interpolation equations in the following table for the number of agreements, which is useful when calculating the Cohen's *Kappa* (K) and Perreault and Leigh's I_r, as well as for the number of weighted coincidences for all the category pairs c-c $\left(o_{cc} = \frac{n_{cc}}{m_u - 1}\right)$ and the sum of the marginal frequencies product of each category ($\sum_c n_c(n_c - 1)$) for Krippendorff's α (see Table IV).

The estimations obtained are represented in the figures below, which provide an overall vision of the behavior of the reliability index for different numbers of categories and judges.

The grading represented in Figure 2 serve as a reference for evaluating the agreement level obtained in a coding process. As we can see, the Cohen's *Kappa* and Perreault and Leigh's I_r show a behavior and gradings that are similar, while Krippendorff's α is clearly different.

On another note, the estimated reliability of Perreault and Leigh's I_r rises as the number of response categories grows, under "diminishing returns". Even so, its values are, in all the cases, higher than Cohen's *Kappa*. Finally, we should underline that the assumption of linearity in the

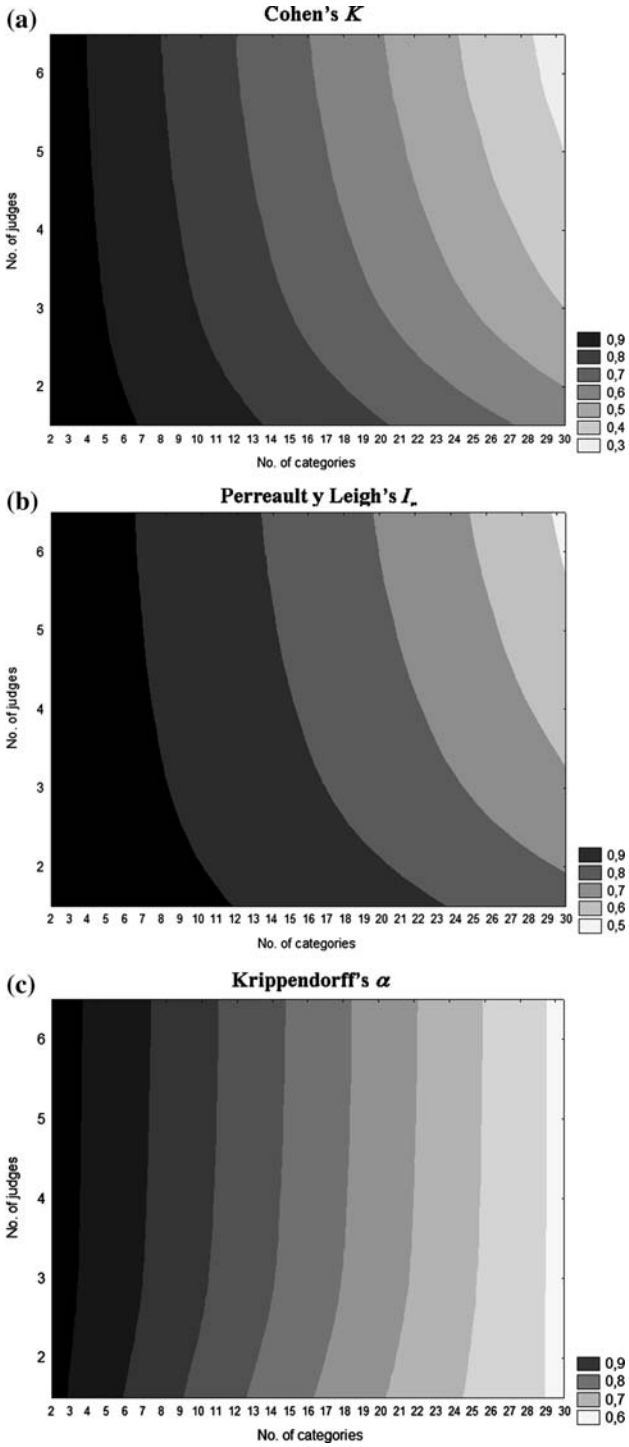

Figure 2. Value of reliability coefficients, in terms of number of judges and categories.

estimation of the number of categories is clearly reflected in the case of the Krippendorff's α coefficient, with a gradual drop.

The above graphs illustrate the inverse relationship between the value of the intercoder reliability and the number of judges and coding categories that are taken into consideration. This shows that, when comparing the level of concordance obtained in different studies that use the same measurement coefficient, we need to compare the magnitudes in the light of this latter aspect. Likewise, the estimations made can be a good starting point in planning the coding process, both in determining the number of judges to be used and the categories to be defined, in order to obtain an adequate reliability.

4. Conclusions and Implications

Interjudge reliability is often perceived as the standard measurement of the coding process' quality. However, high interjudge agreement values could be masking scarce operational deficiencies, as well as errors in establishing categories or in the training of judges. That is, the use of interjudge reliability coefficients is of obvious importance, but if, for example, the judges are making consonant, though incorrect, judgments, the coefficient lacks all meaning (Kolbe and Burnnett, 1991). In this sense, we have observed that the coding of open-ended questions is a laborious process and is subject to errors, both sampling and non-sample. Furthermore, quantitative, descriptive and systematic criteria and procedures are applied and should be accompanied by an exhaustive control and constant revision, which would help guarantee the scientific nature of the technique used.

The variations on the original formulation of Cohen's *Kappa*, with a view to application in marketing research, have come about because the coefficient (a) is conservative, due to the way in which the judgments that coincide by chance are calculated (more specifically, the number of categories has an inverse effect on the probability of obtaining chance agreements); (b) is useful for those situations in which it is expected that few cases will be assigned to certain "strange" categories; and (c) in which it is expected that the set of response patterns will be evaluated by comparing it with a standard defined (Perreault and Leigh, 1989; Brenner and Kliebsch, 1996; Hsu and Fied, 2003).

This study has used real data to model and generalize the behavior of three coefficients of interjudge reliability: Cohen's *Kappa* (K), Krippendorff's α and Perreault and Leigh's I_r, in terms of the number of judges and categories used. The coefficients' generalizations, at any value of these

two characteristics, have been made possible by introducing a series of corrections in their formulation.

The results reveal the presence of an inverse relationship between the value reached by the coefficients and, on the one hand, the number of judges and, on the other, the number of categories. Likewise, and in line with the results obtained by Kang et al. (1993), the Perreault and Leigh coefficient obtains higher values than Cohen's *Kappa*. This leads us to consider that this coefficient shows a better behavior (Rust and Cooil, 1994) for those situations in which there are a high number of categories; a situation that is generally common to most consumer and opinion studies in which open-ended questions are used.

Nonetheless the Krippendorff α coefficient proves to be very stable in measuring the concordance granted by multiple judges. Therefore, given that, on increasing the number of coders, the value obtained by the K and I_r coefficients is significantly penalized, the α coefficient is much more useful for the case of a high number of coders. More specifically, we propose that the α coefficient be used in those situations in which there are more than two judges, given the drawbacks of Perreault and Leigh's I_r.

Using the maps of estimated intercoder reliability, the researcher can, *a priori*, determine the most adequate number of judges and categories so as to reach an objective value for the reliability coefficient. With regard to the above, we should underline that the results obtained in the reliability indices calculated are much conditioned by the high number of categories defined in this empirical study.

The main limitation of this study lies in the interpolation carried out of the agreement proportion for different numbers of categories, under the assumption of linearity. For the case of five categories or less, the behavior of the expected agreement percentage, as a measurement of reliability, has been modeled by Rust and Cooil (1994). To do this, we need to carry out a preliminary testing and measure the agreement percentage.

Future investigations in this field should include the actual behavior of the reliability coefficients in terms of this variable, in order to develop a unifying theoretical frame that can be applied in the coding process of open-ended questions.

Acknowledgements

This study has been carried out using funds from the research project forming part of the Spanish National R&D Plan, in turn financed by the EU FEDER funds (ref. 1FD97-0306).

Appendix

Problems that most affect Society and the individual. Frequencies obtained following coding.

Problem	Percentage	
	As a problem affecting Society(%)	As a problem affecting the individual(%)
Unemployment	66.90	51.52
Terrorism	50.21	22.58
Drug and alcohol abuse	35.95	13.79
Problems concerning loss of social values	28.96	18.94
Problems concerning distribution of wealth	18.69	8.03
Delinquency	15.41	14.09
Problemas concerning natural environment	12.98	8.18
Violence	8.27	4.39
Immigration	7.70	3.18
Economic problems	7.42	8.18
Problems concerning educational system	6.42	6.21
Problems concerning the individual	6.28	4.39
Racism	6.13	3.03
Political problems	4.71	2.27
Problems concerning world peace	4.71	1.67
Problems stemming from modern lifestyles	4.14	4.09
Problems concerning women	3.71	2.42
Problems concerning justice	3.57	3.03
Problems concerning health	3.42	3.79
Problems concerning youth	3.00	3.18
Others (that do not fit in any category)	3.00	3.33
Job precariousness	2.85	3.64
Problems concerning Health System	2.43	3.03
Problems concerning lack/cost of housing	2.00	3.33
Problems concerning livestock farming and foodstuffs	2.00	1.82
Problems concerning old age	1.85	1.82
Problems concerning family	1.71	1.67
Public Services	1.28	2.73
Problems concerning childhood	0.57	1.36

Note

1. As specialised software we can mention TEXTSMART 1.0, ProGAMMA, AGREE 7.0, Atlas.ti, VERBASTAT 3.0 or VERBATIM-BLASTER. Likewise, the potentialities of automatization and data processing offered by certain statistical and spreadsheet software applications are of great use. For a more in-depth overview of specialised software packages in qualitative research and their advantages, see Miles and Haberman (1994), Richards and Richards (1994), Weitzman and Miles (1995), Krippendorff (1997) and Lombard *et al.* (2002).

References

Abernethy, A. M. & Franke, G. R. (1996). The information content of advertising: A meta-analysis. *Journal of Advertising* 25(2): 1–17.

Bardin, L. (1977). *L'analyse de contenu*. (Paris: PUF).

Bennett, E. M. & Goldstein, A. C. (1954). Communications through limited response questioning. *Public Opinion Quaterly* 18: 303–308.

Berelson, B. (1952). *Content Analysis in Communications Research*. Glencoe (Ill.): Free Press.

Brenner, H. & Kliebsch, U. (1996). Dependence of weighted kappa coefficients on the number of categories. *Epidemiology* 7(2): 199–202.

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

Denzin, N. K. & Lincoln, Y. S. (1994). *Handbook of Qualitative Research*. Thousand Oaks (CA): Sage Publications.

ESOMAR (2003). *Annual Study of the Market Research Industry 2002*. (Amsterdam: ESOMAR).

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

Fontana, A. & Frey, J. H. (1994). Interviewing. The art of science. In: N. K. Denzin & Y. S. Licoln (eds.),*Handbook of Qualitative Research*. Thousand Oaks (CA): Sage Publications.

Glaser, B. G. & Strauss, A. L. (1999). *The Discovery of Grounded Theory: Strategies for Qualitative Research*. New York: Aldine de Gruyter.

Grayson, K. & Rust, R. (2001). Interrater reliability. *Journal of Consumer Psychology* 10 (1&2): 71–73.

Herbert, L. (1977). Kappa revisited. *Psychological Bulletin*: 84(2): 289–297.

Holbrook, M. B. (1977). More on content analysis in consumer research. *Journal of Consumer Research*: 4(3): 176–177.

Holsti, O. R. (1969). *Content Analysis for the Social Sciences and Humanities*. Reading (MA): Addison-Wesley.

Hsu & Fied (2003). Interrater agreement measures: Comments on Kappa, Cohen's Kappa, Scott's π and Aicking α. *Understanding statistics* 2(3): 205–219.

Hughes, M. A. & Garret, D. E. (1990). Intercoder reliability estimation approaches in marketing: A Generability Theory framework for quantitative data. *Journal of Marketing Research* 27(2): 185–95.

Kang, N., Kara, A., Laskey, H. A. & Seaton, F. B. (1993). A SAS MACRO for calculating intercoder agreement in content analysis. *Journal of Advertising* 22(2): 17–28.

Kassarjian, H. H. (1977). Content analysis in consumer research. *Journal of Consumer Research* 4(2): 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* 18(September): 243–250.

Krippendorff, K. (1971). Reliability of recording instructions: Multivariate agreement for nominal data. *Behavioral Science* 16(3): 228–235.

Krippendorff, K. (1980). *Content Analysis, an Introduction to Its Methodology*. Thousand Oaks (CA): Sage Publications.

Landis, J. R. & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics* 33: 159–174.

Lehmann, R. L., Gupta, S. & Steckel, J. H. (1998). *Marketing Research*. Reading (MA): Addison-Wesley Educational Publishers Inc.

Light, R. J. (1971). Measures of response agreement for qualitative data: some generalizations and alternatives. *Psychological Bulleting* 76(5): 365–377.

Lincoln, Y. S. & Guba, E. G. (1985). *Naturalistic Inquiry*. Beverly Hills (CA): Sage Publications.

Lombard, M., ZINDER-duch, J. & Campanela, C. (2002). Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communication Research* 28(4): 587–604.

López-Aranguren (1989). El análisis de contenido. In: M. García et al. (Coord.): *El Análisis de la Realidad Social*: *Métodos y Técnicas de Investigación.* Madrid: Alianza Editorial.

Luque, T. (1997). *Investigación de Marketing*. Barcelona: Ed. Ariel.

Mackensen, K. & Wille, U. (1999). Qualitative text analysis supported by conceptual data systems. *Quality & Quantity* 33: 135–156.

Miles, M. B. & Huberman, A. M. (1994). *Quality Data Analysis. An expanded Sourcebook*. Thousand Oaks (CA): Sage Publications.

Miller, W. L. & Crabtree, B. F. (1994). Clinical research. In: N. K. Denzin & Y. S. Licoln (eds.), *Handbook of Qualitative Research*. Thousand Oaks (CA): Sage Publications.

Perreault, W. D. & LEIGH E. L. (1989). Reliability of nominal data based on quantitative judgments. *Journal of Marketing Research* 23(May): 130–43.

Richards & Richards (1994). Using computers in qualitative research. In: N. K. Denzin & Y. S. Licoln (eds.), *Handbook of Qualitative Research*. Thousand Oaks (CA): Sage Publications.

Riffe, D. & Freitag, A. A. (1997). A content analysis of content analysis: Twenty-five years of Journalism Quarterly. *Journalism & Mass Communication Quarterly* 74: 873–882.

Roberts, C. W. (2000). A conceptual framework for quantitative text analysis: On joining probabilities and substantive inferences about text. *Quality & Quantity* 34: 259–274.

Rust, R. T. & Cooil, B. (1994). Reliability measures for qualitative data: Theory and implications. *Journal of Marketing Research* 31: 1–14.

Spiggle, S. (1994). Analysis and interpretation of qualitative data in consumer research. *Journal of Marketing Research* 21 (December): 491–503.

Spitznagel, E. L. & Helzer, J. E. (1985). A proposed solution to the base rate problem in the kappa statistic. *Archives of General Psychiatry* 42(7): 725–28.

Strauss, A. & Corbin, J. (1990). *Basic of qualitative research: Grounded theory procedures and techniques*. Newbury Park (CA): Sage Publications.

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

Weitzman, E. A. & Miles, M. B. (1995). *Computer Programs for Qualitative Data Analysis: A Software Sourcebook*. Thousand Oaks (CA): Sage Publications.