The Effect of Asymmetry on the 2x2 Kappa Coefficient: Application to the Study of Learning Disabilities

Teresa Rivas-Moya[1] and María-José González-Valenzuela
Malaga University, Spain

In educational practice, for the evaluation and diagnosis of learning disabilities (LD), it is advisable to use standardized tests together with observation questionnaires. When observation questionnaires are used in the study of LD, Cohen's (1960) kappa coefficient **(κ)** *is frequently applied as a measure of agreement between two raters when they independently classify a sample of subjects in several categories. In practice, a good interpretation cannot be made if the conditions surrounding the calculation are not taken into consideration. This investigation presents a study of asymmetry and its effect on the κ interpretation. In Study 1, the importance of symmetry is highlighted by means of several examples that show agreement between two raters when classifying 60 subjects in one of two categories. From these examples the interpretation of κ is complemented with the information given by (a) asymmetry analyzed by descriptive and graphical methods and hypothesis tests; and (b) other values, such as maximum observed agreement, maximum reachable agreement, and maximum unreachable agreement. In Study 2, the concepts of Study 1 are applied to examples of LD.*

Key Words: Agreement, Kappa, Symmetry, Learning Disabilities

Among the oldest and most persistent questions in the field of learning difficulties (LD) are its definition and assessment. The definition of LD is a complex task for educators and researchers alike, due in large part to the plurality of its historical roots, perspectives, and theoretical models. The debate surrounding the definition of LD means that its research and assessment must be re-examined, for various reasons. One reason worth noting is the advisability of defining (a) the properties of the measures, methods, and requirements to optimize the diagnostic process; and (b) the type of instruments, strategies, or assessment approach suitable for applying information in the treatment and determining its needs (Jiménez, 1999). Thus, if the models and assessment measures are reliable, they may serve to throw light on the definition of LD and its connection with instruction in an effort to prevent or improve LD. In educational practice, the most suitable assessment model is one that combines a static or standardized assessment with a dynamic or observational assessment (Fuchs & Fuchs, 1996).

The National Joint Committee on Learning Disabilities (NJCLD, 2006) defines LD as a general term referring to a heterogeneous group of disorders mani-

1. Please address correspondence to: Teresa Rivas Moya, Department of Psychobiology and Methodology, Psychology Faculty. Málaga University, Campus de Teatinos s.n. 29071 Málaga; E-mail:moya@uma.es

fested by significant difficulties in the acquisition and use of listening, speaking, reading, writing, reasoning, or mathematical skills. These disorders are intrinsic to the individual, presumably due to a central nervous system dysfunction, which may occur at any time in life. Problems in self-regulatory behaviors, social perception, and social interaction may exist in individuals with LD, but do not by themselves constitute a LD. Although LD may occur concomitantly with other handicapping conditions (e.g., sensory impairment, mental retardation, serious emotional disturbance) or with extrinsic factors such as cultural differences, inappropriate or insufficient educational instruction), they are not the result of such influences or conditions.

This definition agrees with the definitions proposed by other associations, such as NACHC (National Advisory Committee on Handicapped Children), ACLD (Adults and Children with Learning and Developmental Disabilities), ICLD (Interagency Committee on Learning Disabilities), and also with that given in the *Diagnostic and Statistical Manual* (Pichot, Lopez-Ibor, & Valdés, 1995). Besides, it is widely accepted among professionals and researchers in the field of LD. This definition is based on the acquisition of skills (in reading, writing, mathematics, etc.) implicit within a model of assessment centered on abilities and the product (i.e., a static assessment). Static assessment is characterized as being a standard assessment of psychological abilities or diagnostic procedures. This makes it possible to detect individuals with LD and distinguish their condition from other pathologies. However, to a great extent, it is disconnected from educational intervention (Hammill & Larsen, 1978). The work of Shapiro, Buckhalt, and Herod (1995) is an example of static assessment. The authors examined the performance characteristics of school-identified students with LD using the DAS battery (The Differential Ability Scales; Elliot, 1990) by individual measure of aptitude and achievement levels (verbal, space, reasoning, spelling, reading of words, memory, etc., subtests) defined for individuals ranging in age from 2 years 6 months to 17 years 11 months. The study of Reynolds (1998) is also worthy of mention, in which the TOMAL (Test of Memory and Learning; Reynolds & Bigler, 1994) was applied to a sample of adolescents with LD to assess their performance in memory and learning; TOMAL is a standardized test administered to children ages 5-19 years old.

Another definition of LD is upheld by the National Information Center for Children and Youth Disabilities (NICHY, 2000) and the Learning Disabilities Association of America (LDA, 2006). Here LD is defined as a neurological disorder that affects one or more of the basic psychological processes involved in understanding or using of the spoken or written language. The disability may manifest itself in difficulties related to listening, thinking, speaking, reading, writing, spelling, or doing mathematical calculations (LDA, 2006).

This definition is closer to the first definitions favored by some authors, pioneers in the study of LD (Bateman, 1965), and comes closer to an assessment model based on the psycho-educational process (González, 1997); that is, dynamic assessment. Dynamic assessment, less used in the study of LD, is based more on observational assessment (of teachers and/or parents) of the processes involved in children's learning. This type of assessment facilitates differentiation between LD and other disorders. Further, it links the diagnosis to intervention and instruction by facilitating the detection of the needs of students with LD (Kavale & Forness,

1984). Xenitidis, Thornicroft, Leese, and Slade (2000) used the CANDID (The Camberwell Assessment of the Needs of Adults with Development and Intellectual Disabilities) questionnaire to assess the educational needs of adults with LD. The CANDID is used by psychiatrists specializing in mental health, and the kappa coefficient (κ) was used to obtain the reliability between raters. In another study, Raghavan, Marshall, Lockwood, and Duggan (2004) applied the LDCNS (Learning Disabilities Cardinal Needs Schedule) questionnaire to detect the educational needs and the areas affected in a group of subjects with LD. They also use κ to obtain interrater agreement.

Few studies using this type of assessment have focused on the percentage of agreement between two raters when they assign subjects to certain categories (items and/or areas). If the concordance between raters is *z* when they detect subjects with LD or determine their educational needs, this adds one more proof of the reliability of scores to the validation of measures.

Cohen's kappa (1960) is frequently used to analyze the reliability, reproductivity, concordance, or interrater agreement when classifying independently each subject of a sample or group into one of *k* categories.

Table 1 shows a 2x2 table for the classification of n subjects – of a group or sample – given by two raters. The absolute frequencies are noted, $n_{ij}, n_{i}., n._{i}, n$, and the associated proportions should be $p_{ij} = \frac{n_{ij}}{n}, p_{i}. = \frac{n_{i}.}{n}, p._{i} = \frac{n._{i}}{n}, 1$, respectively. The marginal frequency and proportion distributions are given in Table 2.

Table 1
Classification of Two Raters in Two Categories

		Rater B		
		Category 1	Category 2	Total Row
Rater A	Category 1	n_{11}	n_{12}	$n_{1.}$
	Category 2	n_{21}	n_{22}	$n_{2.}$
	Total Column	$n_{.1}$	$n_{.2}$	n

Table 2
Marginal Distributions Associated with a 2x2 Table

	Marginal Frequencies		Marginal Proportions	
Category	*Rater A*	*Rater B*	*Rater A*	*Rater B*
1	$n_{1.}$	$n_{.1}$	$p_{1.}$	$p_{.1}$
2	$n_{2.}$	$n_{.2}$	$p_{2.}$	$p_{.2}$
	n	n	1	1

The κ index is defined as $(P_0 - P_c)/(1 - P_c)$. In 2x2 tables, starting from Table 1, $P_0 = \sum(n_{ii}/n) = \frac{n_{11} + n_{22}}{n}$ is the proportion of observed agreement and $P_C = \sum(n_i.n._i/n^2) = \frac{n_1.n._1}{n^2} + \frac{n_2.n._2}{n^2} = \frac{n_1.n._1 + n_2.n._2}{n^2}$ is a specific definition of the proportion of agreement to be expected by chance. If all the agreement is observed agreement (P_0), then $P_C = 0$ and κ reaches its maximum value, 1. In the worst case, all observed agreement is only agreement to be expected by chance. In other words $P_0 = 0$. Then κ reaches its lowest value $\left[-\frac{P_C}{(1 - P_C)}\right]$. Therefore, $-\frac{P_C}{(1 - P_C)} \le \kappa \le 1$ is the range of values for κ. If κ had a value of 0, $\kappa = 0$, the agreement would be equal to the agreement obtained if the classification of subjects were made by chance. If κ were negative, $-\frac{P_C}{(1 - P_C)} \le \kappa < 0$, the agreement obtained would be lower than the agreement to be expected by chance. If κ were positive, $0 < \kappa \le 1$, the agreement would be greater than the agreement obtained when classifying subjects by chance. Although κ statistic does not follow a specific distribution and its value is easy to calculate, κ should not be applied to just any situation of measure in which it is planned to classify a sample of subjects into a set of categories, and to give the extent to which two raters agree in their judgements. Cohen established the following conditions of application: Categories must be independent, mutually exclusive, and exhaustive; raters must operate independently; and the marginal homogeneity (MH) or symmetry (SYM) in the marginal frequency distributions (MFD) must be satisfied. Cohen also gave a specific proportion of agreement to be expected by chance (called the agreement expected by chance measure; AEC. In the formula of κ it is denoted P_C). SYM and AEC are based on the MFD.

Cohen (1960, p. 42) noted the SYM when defining the maximum value of agreement or maximum value of κ, denoted as κ_M and defined as $\kappa_M = \frac{MaxP_0 - P_C}{1 - P_C}$. He reasoned that κ can only reach the maximum value 1 when the MFD are symmetric. He also pointed out that in any study of reliability, the maximum value κ_M permitted by the MFD can be obtained as a function of maximum observed agreement (denoted $MaxP_0$). $MaxP_0$ can be obtained from the marginal proportions $(p_i., p._i \;\; i:1,2)$ given in Table 2 and, in general, it is defined as $MaxP_0 = \max(\sum p_{ii}) = \sum_{i=1}^{2} \min(p_i., p._i) = \min(p_1., p._1) + \min(p_2., p._2)$.

Under perfect SYM, $p_{1.} = p_{.1}$ and $p_{2.} = p_{.2}$, and then $MaxP_0 = \max(\sum p_{ii}) = p_{1.} + p_{2.} = p_{.1} + p_{.2} = 1$. Thus, substituting $MaxP_0$ by 1 in κ_M, $\kappa_M = \frac{1 - P_C}{1 - P_C} = 1$. Otherwise, if the MFD are asymmetric, $MaxP_0 < (p_{1.} + p_{2.}) = (p_{.1} + p_{.2}) = 1$. Then $MaxP_0 < 1$. Also, in κ_M, by substituting $MaxP_0$ with its value <1, $\kappa_M < 1$. Therefore, if the MFD are ASYM, κ cannot reach its maximum value of 1. In such a case, given MFD, Cohen recommended interpreting κ with regard to (a) maximum agreement ($MaxP_0$ and κ_M) that it is possible to obtain and (b) the agreement that it is not possible to obtain ($1 - \kappa_M$). As defined above, given MFD, κ_M is the value of κ calculated when P_0 is the highest possible value. In this case, κ_M is the largest value that κ can reach, and $1 - \kappa_M$ represents the proportion of agreement (chance excluded) that cannot be obtained as a consequence of ASYM. In this way, given MFD, when there is ASYM, $MaxP_0$, κ_M represents the maximum reachable agreement for P_0 and κ, respectively, and $1 - \kappa_M$ the maximum unreachable agreement for κ. In this sense, they complement the interpretation of κ. If these concepts are not considered, it is possible that when there is ASYM, given different MFD and the same P_0, different values of κ would represent the same observed agreement P_0; while κ would be interpreted with respect to the maximum value 1. Given MFD, if this maximum value 1 is impossible to reach, then, an erroneous interpretation of κ would be made.

Several authors have studied, from different perspectives, theoretical and practical questions arising from the use of κ. Some authors have defined agreement measures (AM) alternative to κ, in order to correct the influence of the ASYM and/or AEC. For example, Zwick (1988) analyzed the equivalence between some indices proposed as S (Bennett, Alpert , & Goldstein, 1954), C (Janson & Vegelius, 1979), κ_k (Brennan & Prediger, 1981); and in the case of only two categories, G (Holley & Guilford, 1964) and random error (Maxwell, 1977). The coefficients κ, π (Scott, 1955) and S (Bennett et al., 1954), are similar, but they differ in their AEC and their approach to the SYM. Zwick (1988) also suggested—as did Maxwell (1970) and Fleiss and Everitt (1971) — that the SYM between raters should be analyzed. Brennan and Prediger (1981) studied the implications of some conditions of κ as an index of reliability: the AEC, its maximum value, and the use of a priori non fixed MFD. Gwet (2001) defined the $AC1$ similar to k, with the AEC being $P_C = 2p_{1.}(1 - p_{1.})$. Martín-Andrés and Femia-Marzo (2004) developed the Delta (Δ) based on (a) κ being very influenced by the MFD, (b) a probabilistic model, and (c) the formula for evaluating a multiple-choice test without penalizing the incorrect answers of Hutchinson (1982). So far, none of these AM has replaced the extent of κ's applicability.

Uebersax (2003a) quoted references in which different authors have dealt with the limitations encountered in the application of κ. He made a critical revision of the κ coefficients, describing the pros and cons. He also described statistical methods to analyze the SYM and the agreement between raters for different types of categories (nominal, ordinal, interval).

Given the necessity to evaluate agreement between raters in LD diagnosis, and the fact that κ is one of the more frequently used coefficients, a method to analyze the conditions in which κ can be applied is shown. Based on the works of Cohen (1960), Maxwell (1970), Everitt (1977), Bishop, Fienberg, and Holland (1975), Agresti (1990), Uebersax (2003b, 2003c), and Rivas (2005), this paper proposes a way to analyze the assumption of SYM, and also to calculate and interpret κ. To this end, in Study 1 (a) a detailed study of SYM and its causes should be made, on a descriptive or an inferential level (but only when this makes sense); and (b) from a given MFD, interpretation of κ should be complemented with the information of the values such as $MaxP_0$ that can be reached, and other values related to $MaxP_0$, such as κ_M and $1-\kappa_M$. In Study 2, in the area of LD, examples of application are shown.

STUDY 1

AEC and ASYM are factors that influence κ. If there is SYM, then AEC influences κ. Sometimes the effects of ASYM and AEC become confused when analyzing their influence over κ. Examples of this can be found in which a high observed agreement is unequally distributed in the two categories, and a low observed disagreement is equally (or unequally) distributed in the two categories. In such cases, it may be difficult to justify the need to study the SYM. From the following examples, it will be shown that the ASYM can influence κ, even though the ASYM has no influence (or hardly any) on the AEC. However, study of the ASYM is necessary when calculating κ, whether or not it has any influence over AEC. Another question, not studied here, is how ASYM influences AEC, given that AEC is also calculated from MFD.

The following examples show a high observed agreement equally distributed in the two categories, and a low observed disagreement unequally distributed in the two categories. Thus, in examples of the following type, it may be seen that the same (low) number of disagreements, unequally distributed over the two categories, have a varying influence on the ASYM and κ values. In addition, they hardly influence the AEC.

Given the variety of information contained in the κ value, this study sought to draw researchers' attention to the fact that when applying κ the same equally distributed observed agreement and similar AEC can give similar κ values. However, interpretation of agreement can be different due to the ASYM of MFD.

Method

Given a kxk table, let $n_{ij}\ (i,j:1,...,k)$ be the observed frequencies. If the expected values $N_{ij}(i,j:1,...,k)$ satisfy the MH, then $N_{i\cdot}=N_{\cdot i}\ (i:1,...,k)$. By SYM, Bishop et al. (1975, pp. 281-282) meant that $N_{ij}=N_{ji}\ (\text{for all } i\neq j)$. Similarly, if $p_{ij}\ i,j:1,...,k$ are the observed proportions, and their respective expected

values are $P_{ij}\ (i,j:1,...,k)$, then by SYM, Everitt (1977, p. 114) meant that $P_{ij} = P_{ji}\ (i \neq j)$, and by MH $P_{i\cdot} = P_{\cdot i}\ (for\ i:1,...,k)$. This definition of SYM implies MH. For a 2x2 table, $P_{ij}, P_{i\cdot}, P_{i}\ (i:1,2)$ are the marginal proportions in the population and $p_{ij}, p_{i\cdot}, p_{\cdot i}\ (i:1,2)$ are the observed proportions (see Table 2). If $P_{1\cdot} = P_{\cdot 1}$ and $P_{2\cdot} = P_{\cdot 2}$, this implies that $P_{12} = P_{21}$. In 2x2 tables, the concept of MH is equivalent to the concept of SYM. Thus, in 2x2 tables, reference will be made to SYM. There will be asymmetry (ASYM) if the SYM is not satisfied.

The SYM—or equality of proportions of classification given by two raters in both categories—is tested by the null hypothesis

$H_0 : P_{i\cdot}(Rater\ A) = P_{\cdot i}(Rater\ B)\ \ for\ i:1,2$, or similarly

$H_0 : P_{i\cdot}(Rater\ A) - P_{\cdot i}(Rater\ B) = 0\ \ for\ \ i:1,2$

as the difference $d = (P_{2\cdot} - P_{\cdot 2}) = (1 - P_{1\cdot}) - (1 - P_{\cdot 1}) = (P_{1\cdot} - P_{\cdot 1})$, it is then only necessary to test $H_0 : P_{1\cdot}(Rater\ A) = P_{\cdot 1}(Rater\ B)$, but for tables 2x2 $(P_{1\cdot} - P_{\cdot 1}) = (P_{12} - P_{21})$ (Agresti, 1990; p. 348) then $H_0 : P_{12} = P_{21}$ or $H_0 : P_{12} - P_{21} = 0$.

The McNemar test can be applied to test this hypothesis, whose statistic is:

2.1. $(n_{12} - n_{21})^2 / (n_{12} + n_{21})$, which, under the null hypothesis, is distributed as a χ_1^2 distribution (Maxwell, 1970, p. 653), or

2.2. $(|n_{12} - n_{21}| - 1)^2 / (n_{12} + n_{21})$ using the χ^2 approximation with correction (Bishop et al.,1975; p. 258)

2.3. $Z_U = (2n_{12} - 1 - n)/\sqrt{n}$ and $Z_L = (2n_{12} + 1 - n)/\sqrt{n}$ are large-sample $(n \geq 100)$ test statistics. A significant result is obtained if $Z_U \geq z^U_{\alpha/2}$ or $Z_L \leq z^L_{\alpha/2}$. $z^U_{\alpha/2}$ and $z^L_{\alpha/2}$ are, respectively, the upper and lower critical values of the standard normal distribution. The test is significant if $p'_U \leq \alpha/2$ or $p'_L \leq \alpha/2$ (Krauth, 1990; pp. 113-114).

Given a 2x2 table, test 2.1 can be obtained with the Uebersax (2000) free program, test 2.2 with the Statistica Program, and test 2.3 can be obtained manually using a standard normal distribution table.

If the null hypothesis is rejected in any of the above tests, $P_{12} \neq P_{21}$ can be concluded; therefore, the ASYM comes from the population from which the sample was drawn.

Before interpreting κ, these concepts can be used,

1. If n is a given group, one can make a descriptive study of SYM. In such a case, it is not possible to carry out tests of hypothesis on the SYM. Instead, it is proposed to compare the proportions and see if they are numerically equal to declare SYM.
2. If n is a random sample of a population, an inferential study of SYM can be made by tests of hypotheses on equality of proportions, described in (2.1) (2.2) and (2.3) above.

SYM can be analyzed on a descriptive or inferential level, and the interpretation of κ should be given based on the results of the SYM study made on a descriptive level. If SYM is assumed on a descriptive level, κ is interpreted with regard to maximum value 1. If SYM cannot be assumed, κ is interpreted in respect of $MaxP_0, \kappa_M, 1-\kappa_M$ values. This is applied in the following examples.

In Examples 1-5 (Ex. 1-5), two raters classify 60 subjects in one of two categories. The total number of agreements between raters $n_{11}+n_{22}$ (50) is distributed equally over both categories. The total number of disagreements $n_{12}+n_{21}$ (10) is distributed differently (5, 5) (6, 4) (8, 2) (9, 1) (10, 0) in the two cells associated with the disagreements (see Table 3).

Table 3
2x2 Tables Examples 1-5

Example 1				
		Rater B		
		Category 1	Category 2	Total Row
Rater A	Category 1	25	5	30
	Category 2	5	25	30
	Total Column	30	30	60
Example 2				
		Rater B		
		Category 1	Category 2	Total Row
Rater A	Category 1	25	4	29
	Category 2	6	25	31
	Total Column	31	29	60
Example 3				
		Rater B		
		Category 1	Category 2	Total Row
Rater A	Category 1	25	2	27
	Category 2	8	25	33
	Total Column	33	27	60
Example 4				
		Rater B		
		Category 1	Category 2	Total Row
Rater A	Category 1	25	1	26
	Category 2	9	25	34
	Total Column	34	26	60
Example 5				
		Rater B		
		Category 1	Category 2	Total Row
Rater A	Category 1	25	0	25
	Category 2	10	25	35
	Total Column	35	25	60

From the data in Table 3, it can be seen that the observed agreement does not influence the ASYM, the observed disagreement influences the ASYM, and the ASYM hardly influences the different AEC values.

Results

Results of the SYM study are shown in Table 4. Columns 3-4 and 5-6 show the absolute frequency and marginal proportion distributions. Distributions of marginal proportions (Columns 5-6) are shown in Figure 1.

Table 4
Marginal Proportion and Frequency Distributions and McNemar Test

Example	Category	Rater A	Rater B	Rater A	Rater B	χ_1^2 (p)
		Frequency		Proportion		
1						
	1	30	30	0.500	0.500	0.00(1.000)
	2	30	30	0.500	0.500	
2						
	1	29	31	0.483	0.517	0.10(0.752)
	2	31	29	0.517	0.483	
3						
	1	27	33	0.450	0.550	3.60(0.058)
	2	33	27	0.550	0.450	
4						
	1	26	34	0.433	0.567	6.4(0.011)
	2	34	26	0.567	0.433	
5						
	1	25	35	0.417	0.583	10.0(0.002)
	2	35	25	0.583	0.417	

On a descriptive level, if ASYM is considered when there is any difference between $p_{i.}$ and $p_{.i}$ $(i{:}1,2)$, SYM is assumed in Ex.1 and ASYM is assumed in Ex. 2, 3, 4, 5.

On an inferential level, if it is assumed that making educated guesses – say a significance level of 0.05 – this would give an estimated random sample size $n = 60$. So the hypothesis of SYM could then be tested with the McNemar test (see 2.1). The statistical significance could then be interpreted (Column 7, Table 4).

There are no significant differences between raters when they classify subjects in both categories in Ex. 1 (p= 1.000), Ex. 2 (p = 0.752) and Ex. 3 (p = 0.058). In Ex. 1-3, the hypothesis of SYM is not rejected.

There are significant differences between raters when they classify subjects in both categories in Ex. 4 (p<0.05) and Ex. 5 (p<0.01). In Ex. 4-5 the hypothesis of SYM is rejected, so ASYM of MFD can be considered.

If there is ASYM, the causes should be studied, after which κ and related values may be analyzed.

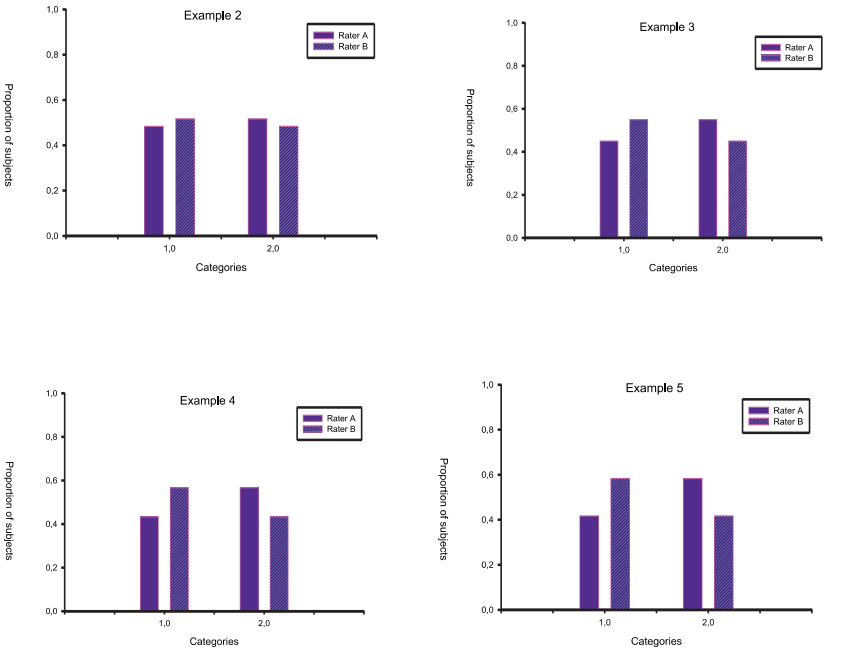


Figure 1. Histograms of marginal proportion distributions (examples 2-5).

Table 5 shows the values of observed agreement (P_0), agreement to be expected by chance (P_C), kappa (κ), maximum of P_0 given the MFD $(MaxP_0)$, maximum value of κ obtained in $MaxP_0$ (κ_M), and $(1-\kappa_M)$. The values of the indices given in Table 5 are shown in Figure 2.

Table 5
κ and Related Values

Examples	P_0	P_C	κ	$MaxP_0$	κ_M	$1-\kappa_M$
1	0.833	0.500	0.666	1.000	1.000	0.000
2	0.833	0.499	0.667	0.967	0.941	0.059
3	0.833	0.495	0.669	0.900	0.808	0.192
4	0.833	0.491	0.672	0.867	0.739	0.261
5	0.833	0.486	0.675	0.833	0.675	0.325

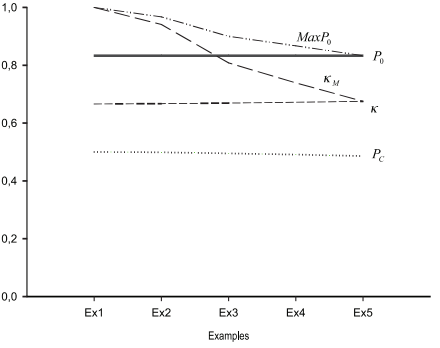


Figure 2. Plot of K and related values.

From the results of Table 5, the following interpretation can be made.

On a descriptive level, Ex. 1 SYM is observed. The agreement is $\kappa = 0.666$ with regard to $\kappa_M = 1$, $MaxP_0 = 1$ is reachable, and $P_C = 0.5$. Because of SYM, all possible agreement is reachable $(1 - \kappa_M = 0)$.

Ex. 2 ASYM is observed. The agreement is $\kappa = 0.667$ with regard to $\kappa_M = 0.941$, $MaxP_0 = 0.967$ is reachable, and $P_C = 0.499$. There is hardly any unreachable agreement because of the ASYM $(1 - \kappa_M = 0.059)$.

A similar interpretation applies to Ex. 3 and Ex. 4. The maximum reachable agreement decreases and maximum unreachable agreement increases until Ex. 5.

In Ex. 4, the agreement is $\kappa = 0.672$ with regard to $\kappa_M = 0.739$, $MaxP_0 = 0.867$ is reachable, and $P_C = 0.491$. However, it is not possible to reach a meaningful degree of agreement owing to ASYM $(1 - \kappa_M = 0.261)$.

Ex. 5 ASYM is also observed. The agreement is $\kappa = 0.675$ with regard to $\kappa_M = 0.675$, $MaxP_0 = 0.833$ is reachable, and $P_C = 0.486$. However, it is not possible to reach a meaningful degree of agreement owing to ASYM $(1 - \kappa_M = 0.325)$.

In summary, given P_0 in each 2x2 table, the greater the ASYM, the lower the obtainable maxima $MaxP_0$ and κ_M. Therefore, the greater the ASYM, the greater the unobtainable maximum, $1 - \kappa_M$.

In Table 5 and Figure 2, all the examples show an observed agreement of 0.833 (Column 2). The AEC is close or equal to 0.50 (Column 3). It hardly decreases when the degree of ASYM increases. κ is a value around 0.67 (Column 4). Given a MFD, the maximum values that P_0 $(MaxP_0)$ can reach range from 1 (SYM – Ex. 1) to the observed agreement 0.833 (the greatest degree of ASYM, when a cell of disagreement is zero – Ex. 5) (Column 5). The maximum values that κ can reach, κ_M, go from 1 (SYM – Ex. 1) to 0.675 (the greatest degree of ASYM – Ex. 5) (Column 6).

STUDY 2

Study 2 applies the method and the analysis used in Study 1 to supposed data obtained from assessment of subjects with learning disabilities (*LD*) and without ($\overline{LD}$). Two examples are proposed in which the reliability between pairs of raters can be analyzed.

Let us suppose n = 350 subjects (aged 7-8), in second grade. They are drawn from different state schools in lower sociocultural areas. An educational psychologist selected children of normal intellectual level and without physical, mental or sensory disabilities. This psychologist (Rater A), together with a qualified special education teacher (Rater B) and general educator (Rater C), assessees *LD* or $\overline{LD}$ subjects.

Case 1

A given group (n = 350), nonrandomly drawn from a population of children (second grade, aged 7-8, from different state schools in lower sociocultural areas). The educational psychologist identifies 32 *LD* subjects. An equal number of $\overline{LD}$ subjects (32), matched on age and sex with each *LD* subject, is drawn from the group of 318 $\overline{LD}$ children. The 64 subjects identified by Rater A, 32 *LD* and 32 $\overline{LD}$ subjects, are also independently assessed by Raters B and C.

Table 6 shows 2x2 tables of classification of the 64 subjects, in categories *LD* or $\overline{LD}$, given by the pairs of raters (A, B) (A, C) (B, C).

Table 6
2x2 Tables for Raters (A ,B) (A, C) (B, C)

		Rater A		
		LD	$\overline{LD}$	Total Row
Rater B	LD	31	6	37
	$\overline{LD}$	1	26	27
	Total Column	32	32	64

		Rater A		
		LD	$\overline{LD}$	Total Row
Rater C	LD	31	12	43
	$\overline{LD}$	1	20	21
	Total Column	32	32	64

		Rater B		
		LD	$\overline{LD}$	Total Row
Rater C	LD	35	8	43
	$\overline{LD}$	2	19	21
	Total Column	37	27	64

SYM is analyzed on a descriptive level. There is ASYM in tables (A, B) (A, C) (B, C) (Col 2-3, Col 4-5 in Table 7). Results of agreement are given in Table 8 and shown in Figure 3.

Table 7

Marginal Proportion and Frequency Distributions and McNemar Test for Raters (A, B) (A, C) (B, C)

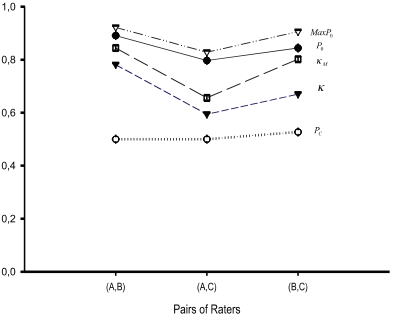
Category	Frequency		Proportion	
	Rater A	*Rater B*	*Rater A*	*Rater B*
LD	32	37	0.500	0.578
$\overline{LD}$	32	27	0.500	0.422
	Rater A	*Rater C*	*Rater A*	*Rater C*
LD	32	43	0.500	0.672
$\overline{LD}$	32	21	0.500	0.328
	Rater B	*Rater C*	*Rater B*	*Rater C*
LD	37	43	0.578	0.672
$\overline{LD}$	27	21	0.422	0.328

Table 8

κ and Related Values for Raters (A, B) (A, C) (B, C)

Rater	P_0	P_C	κ	$\kappa_{\min}$	$MaxP_0$	κ_M	$1-\kappa_M$
A,B	0.891	0.500	0.781	-1.000	0.922	0.844	0.156
A,C	0.797	0.500	0.594	-1.000	0.828	0.656	0.344
B,C	0.844	0.527	0.670	-1.114	0.906	0.802	0.198

Figure 3. Plot of κ and related values for raters (A, B), (A, C), (B, C).



Observed agreement, P_0, is hardly greater (0.891) for Raters A, B than for Raters B, C (0.844). For Raters A, C, observed agreement (0.797) is lower than for

(A, B) and (B, C). Given the respective MFD of rater pairs, the $MaxP_0$ reachable are 0.922, 0.828 and 0.906 for (A, B), (A, C) and (B, C), respectively. P_C is around 0.5. P_C is hardly greater (0.527) for (B, C), where ASYM in observed agreement (35, 19) is greater than (31, 26) for (A, B) and (31, 20) for (A, C) (see Table 6).

In (A, B), $\kappa = 0.781$ is the degree of agreement between A and B, above what may be expected if subjects are classified by chance. The maximum reachable value is $\kappa_M = 0.844$, and the unreachable agreement due to ASYM is $1 - \kappa_M = 0.156$.

In (A, C), $\kappa = 0.594$ is the degree of agreement between A and C, above what may be expected if subjects are classified by chance. The maximum reachable value is $\kappa_M = 0.656$, and the unreachable agreement due to ASYM is $1 - \kappa_M = 0{,}344$.

In (B, C), $\kappa = 0.670$ is the degree of agreement between B and C, above what may be expected if subjects are classified by chance. The maximum reachable value is $\kappa_M = 0.802$, and the unreachable agreement due to ASYM is $1 - \kappa_M = 0{,}198$. As a consequence of ASYM, the agreement that it is not possible to reach owing to the global classification of each rater (MFDs) is greater in (A, C) than in (B, C) and (A, B), This causes the maximum reachable agreement $(\kappa_M = 0.656)$ in (A, C) to be lower than $(\kappa_M = 0.844)$ and $(\kappa_M = 0.802)$, respectively, in (A, B) and (B, C).

Given this situation, the agreement obtained is

(1) (A, B), $\kappa = 0.781$ with regard to $(\kappa_M = 0.844)$,

(2) (A, C), $\kappa = 0.594$ with regard to $(\kappa_M = 0.656)$,

(3) (B, C), $\kappa = 0.670$ with regard to $(\kappa_M = 0.802)$.

A comparison between (A, C) and (B, C) shows a loss of agreement greater in (A, C) than in (B, C). It can be seen that there are differences in the observed agreements in both cases, though not important enough to justify the decrease of κ. These differences can be due to ASYM, being greater in (B, C), in which case the cause of ASYM in (B, C) should be investigated.

A similar interpretation could be made by comparison between (A, B) and (A, C) or (A, B) and (B, C).

Case 2

In another case, if n' = 350 is the estimated sample size under simple random sampling, and the 350 subjects drawn from the population of children (second grade, aged 7-8, from different state schools in lower sociocultural areas). $n' = 350$ should have been calculated by the following equations:

1) $n' \geq \dfrac{z^2\left(1-\hat{P}\right)}{\varepsilon^2 \hat{P}}$ (approximate)

2) $n' \geq \dfrac{z^2 N\hat{P}\left(1-\hat{P}\right)}{(N-1)\varepsilon^2\hat{P}^2 + z^2\hat{P}\left(1-\hat{P}\right)}$ (exact; N is the population size)

For this, from an a priori study, it would be necessary to know:

1) $\hat{P}$ or the estimated proportion of LD subject,
and ascertain
2) the significance level (e.g. $\alpha = 0.05$ or confidence level 95%, and $z = 1.96$ associated with $(1-\alpha)$ in Normal distribution), and
3) the low acceptable error (e.g. $\left|\hat{P} - P\right| = \varepsilon = 0.001$).

In this random sample, Rater A identifies *LD* and $\overline{LD}$ subjects, who are also independently assessed by Rater B and Rater C. The study of SYM on a descriptive and inferential level, similar to the above examples, could be made. Only a study of agreement in the sample is recommended.

Discussion

The five examples given in Study 1 have made it possible to show that the same number of agreements (50) distributed equally in both categories, and the disagreements (10) distributed differently, give similar κ values, whereas the study of the SYM gives different results (to be borne in mind in the agreement interpretation). Under the same conditions, other similar examples would give results similar to those presented here.

In general, given a large number of agreements distributed equally in both categories similar κ values (≈ 0.7) and similar AEC (≈ 0.5) are obtained. However, under the same conditions, but bearing in mind the asymmetry in the distribution of the disagreements, different maximum values of agreement between raters, κ_M, are obtained. Lower values of maximum agreement κ_M (κ_M obtained in relation to the maximum value that P_0 can reach given the MFD) are obtained when there is a greater degree of ASYM. Similarly, κ_M decreases when the disagreements tend to be concentrated in one of the two cells (or when the ASYM increases). Thus, results of the ASYM and κ_M yield important information that complements the information given by κ.

The interpretation of the measure AEC cannot be obviated due to its meaningful influence on κ. The previous examples show that high values of AEC (≈ 0.5) in relation to the obtained κ values (≈ 0.7) have been obtained. The AEC of Gwet (2001) has less influence on AC1 than the AEC of Cohen on κ. The influence of the AEC on the AM should be the subject of further study because there is no single concept of AEC.

Negative values of κ can be obtained when observed agreement is lower than agreement expected by chance $(P_0 < P_C)$. $\kappa = 0$ when $P_0 = P_C$ which, in practice, should happen rarely, if ever. Then, one would agree with Agresti (1990, p. 366-367) that testing $H_0 : \kappa = 0$ is not important. This author proposes substituting this test of hypothesis with a confidence interval for κ. For multinomial sampling, the sample measure $\hat{\kappa}$ has a large-sample normal distribution and Agresti (1990) gives its estimated asymptotic variance. In other situations (no multinomial sampling and/or no large sample), this confidence interval should be not estimated. Other authors have been able to give other tests of hypothesis or confidence inter-

vals for K, whose conditions (type of sampling, sample distribution and sample size) must be considered when they are applied.

In addition, the inferential study of SYM, as in all tests of hypothesis, is not easy because the required sample size must be calculated. Before calculating the sample size, a confidence level, from which one can generalize, must be chosen. It is also necessary to make educated guesses, so as to establish the maximum difference allowed between the sample estimate of proportion, $\hat{P}$, and the true unknown population parameter, P. In addition, simple random sampling is not always possible due to (a) the cost of implementing it (e.g., in situations where interviews are required), or (b) when required characteristics must be estimated from strata that comprise a small proportion of the total population. In these cases, other types of sampling should be chosen. In this respect, the interested researcher may consult Levy and Lemeshow (1991) or other texts on sampling. If the characteristics of study do not allow drawing a suitable random sample from a population (thereby making the generalization from the sample to the population impossible), one is limited only to describing the results obtained in the given group.

Generalization of ASYM study to kxk tables can be made by use of the appropriate tests of hypothesis and their associated statistics. Although Cohen (1960) put no restriction on the K application to kxk tables (k>2), care should be taken when considering more than two categories. We agree with Maclure and Willett (1987) and Martín-Andrés and Femia-Marzo (2005) that the use of several K from various 2x2 tables obtained from a kxk table can be more informative than only one K obtained from the kxk table.

Agreement is a particular measure of association, and simplifying the degree of agreement into a single K value is not recommended. This agreement value should be interpreted together with all the measures involved in calculating the index (SYM, or more frequently ASYM, and AEC). In addition, careful study of the causes of the ASYM can give us a better insight into the reasons for the asymmetry in the disagreements and/or agreements (categories of classification wrongly established, errors in the instrument of measure used for classification, incongruence between raters when assigning subjects to categories, prevalence of disorder when raters assess agreement between diagnoses, bias, etc.). Tests of hypothesis and confidence intervals for K, as with other statistical indices, should be planned together with the collection of data (if the study allows the definition of a population, sample distribution, size of sample, significance level, etc.).

Evaluation of individuals with LD is a complex task because it is necessary to specify the properties of the different types of instruments, measures, methods, and diagnostic approaches to detect their characteristics and educational needs. In educational practice, the most appropriate model of assessment is one that combines a static evaluation with a dynamic evaluation; that is, an evaluation based on the use of standardized tests and one based on the observations given by raters (psychologists and teachers). This last type of evaluation is less frequent in the study of LD. However, if this evaluation is used and the coefficient K is applied, an accurate interpretation of K cannot be made without taking into consideration the SYM and its consequences. As shown in the examples their in Study 2.

References

Adults and Children with Learning and Developmental Disabilities. (2006). Retrieved March, 2006, from www.acldd.org

Agresti, A. (1990). *Categorical data analysis.* New York: John Wiley & Sons

Bateman, B. (1965). An educator´s view of a diagnostic approach to learning disorders. In F. Hellmount (Ed.), *Learning disorders* (pp. 219-239). Seattle, WA: Special Child Publications.

Bennett, E. M., Alpert, R., & Goldstein, A. C. (1954). Communications through limited response questioning. *Public Opinion Quarterly, 18*, 303-308.

Bishop, Y.M.M., Fienberg, S. E., & Holland P. W. (1975). *Discrete multivariate analysis: Theory and practice.* Cambridge, MA: MIT Press.

Brennan, R. L., & Prediger, D. J. (1981). Coefficient kappa: Some uses, misuses and alternatives. *Educational and Psychological Measurement, 4,* 687-699.

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

Elliot, C. D. (1990). *Differential ability scales: Introductory and technical handbook.* San Antonio, TX: Psychological Corporation.

Everitt, B. S. (1977). *The analysis of contingency tables.* London: Chapman & Hall

Fleiss, J. L., & Everitt, B.S. (1971). Comparing the marginal totals of square contingency tables. *British Journal of Mathematical and Statistical Psychology*, *24,* 117-123.

Fuchs, L.S., & Fuchs. S. D. (1996). Combining performance assessment and curriculum based measurement to strengthen instructional planning. *Learning Disabilities Research & Practice, 11*, 183-192.

González, M. J. (1997). *Las dificultades de aprendizaje desde una perspectiva psicoeducativa.* Málaga, Spain: Servicios de publicaciones de la Universidad de Málaga.

Gwet, K. (2001). *Handbook of inter-rater reliability.* Gaithersburg, MD: Stataxis Publishing Company.

Hammill, D. D., & Larsen, S. C. (1978). The effectiveness of psycholinguistic training. A reaffirmation of position. *Exceptional Children*, *44,* 402-414.

Holley, W., & Guilford, J. P. (1964). A note on the G-index of agreement. *Educational and Psychological Measurement, 24,*749-753.

Hutchinson, T. P. (1982). Some theories of performance in multiple-choice test, and their implications for variants of the task. *British Journal of Mathematical and Statistical Psychology, 35,* 71-89.

Interagency Committee on Learning Disabilities. (2006). *Definition.* Retrieved March, 2006 from www.kidsource.com

Janson, S., & Vegelius, J. (1979). On generalization of the G index and the phi coefficient to nominal scales. *Multivariate Behavioral Research, 14,* 255-269.

Jimenez, J. E. (1999). *Psicología de las dificultades de aprendizaje.* Madrid,Spain: Síntesis.

Kavale, K. A., & Forness, S. R. (1984). A meta-analysis assessing the validity of Wechsler Scale profiles and recategorization: Patterns or parodies? *Learning Disability Quarterly*, *7*, 136-156.

Krauth, J. (1990). *Distribution-free statistics: An application-oriented approach.* Oxford, UK: Elsevier.

Learning Disabilities Association of America. (2006). *Learning disabilities: signs, symptoms and strategies.* Retrieved April 24, 2006, from http://www.ldaamerica.org

Levy, P. S., & Lemeshow, S. (1991). *Sampling of populations: Methods and applications.* New York: Wiley and Sons.

Maclure, M., & Willett, W. (1987). Misinterpretation and misuse of the statistic kappa. *American Journal of Epidemiology, 126,* 161-169.

Martín-Andrés, A., & Femia-Marzo, P. (2004). Delta: A new measure of agreement between two raters. *British Journal of Mathematical and Statistical Psychology*, *57*, 1-19.

Martín-Andrés, A., & Femia-Marzo, P. (2005). Chance-corrected measures of reliability and validity in KxK tables. *Statistical Methods in Medical Research, 14,* 473-492.

Maxwell A. E. (1970). Comparing the classification of subjects by two independent judges. *British Journal of Psychiatry*, *116*, 651-655.
Maxwell A. E. (1977). Coefficients of agreement between observers and their interpretation. *British Journal of Psychiatry, 130*, 79-83.
National Advisory Committee on Handicapped Children (2006). Retrieved March, 2006, from www.nach.com
Pichot, P., Lopez-Ibor, J. J., Valdés, M. (1995). *DSM-IV. Criterios diagnósticos*. Barcelona, Spain: Masson.
Raghavan, R., Marshall, M., Lockwood, A., & Duggan, L. (2004). Assessing the needs of people with learning disabilities and mental illness: Development of the learning disability version of the Cardinal Needs Schedule. *Journal of Intellectual Disability Research*, *48*, 25-36.
Reynolds, C. R. (1998). Reliability of performance on the test of memory and learning (TOMAL) by an adolescent learning disability sample. *Educational and Psychological Measurement, 58*(5), 832-835.
Reynolds, C. R., & Bigler, E. D. (1994). *Test of memory and learning*. Austin, TX: Pro-Ed.
Rivas, T. (2005, September). *The 2x2 kappa coefficient and the condition of symmetry marginal distributions*. Paper presented at the 8th European Conference on Psychological Assessment, Budapest.
Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly, 19,* 321-325.
Shapiro. S. K., Buckhalt, J. A., & Herod, L. A. (1995). Evaluating of learning-disabled students with the differential ability scales (DAS). *Journal of School Psychology*, *33*(3), 247-263.
The National Information Center for Children and Youths with Disabilities. (2000). *Learning disabilities.* Retrieved March, 1, 2000, from http://www.NICHCY.org
The National Joint Committee on Learning Disabilities. (2006). *What is a learning disability?* Retrieved April, 25, 2006, from http://www.ldonline.org/
Uebersax, J. (2000). *MH program.* Retrieved May 26, 2004, from http://ourworld.compuserve.com/homepages/jsuebersax
Uebersax, J. (2003a). Kappa coefficients. In J. Uebersax (Ed.), *Statistical methods for rater agreement* (pp. 1:10, 10:10). Retrieved May 26, 2004, from http://ourworld.compuserve.com/homepages/jsuebersax/agree.htm
Uebersax, J. (2003b). Tests of marginal homogeneity. In J. Uebersax (Ed.), *Statistical methods for rater agreement* (pp. 1:6, 6:6). Retrieved May 26, 2004, from http://ourworld.compuserve.com/homepages/jsuebersax/agree.htm
Uebersax, J. (2003c). McNemar Tests of marginal homogeneity. In J. Uebersax (Ed.), *Statistical methods for rater agreement* (pp. 1:8, 8:8). Retrieved May 26, 2004, from http://ourworld.compuserve.com/homepages/jsuebersax/agree.htm
Xenitidis, K., Thornicroft, G., Leese, M., & Slade, M. (2000). Reliability and validity of the CANDID – A needs assessment instrument for adults with learning disabilities and mental health problems. *The British Journal of Psychiatry*, *176*, 473-478.
Zwick, R. (1988). Another look at interrater agreement. *Psychological Bulletin*, *103*(3), 374-378.

Received December 3, 2006
Revised February 2, 2007
Accepted February 3, 2007

Author's Note

This research was partially funded by grants from the Ministerio de Ciencia y Tecnología. (Project Ref:BSO2001-1945) and Consejería de Educación y Ciencia de la Junta de Andalucía (Research Group CTS-278)