KM 9(1983)
pag 96-112

COMPUTER PROGRAMS ON NOMINAL SCALE AGREEMENT

Roel Popping

Summary

There are several computer programs on nominal scale agreement. One of these is the program AGREE3, which was developed by the author. In part II of this paper a number of computer programs will mainly be compared on the following characteristics: 1) indices that can be computed within a program; 2) input and data matrices; and 3) program languages.

In part I a short theoretical introduction will be given to nominal scale agreement. This introduction is meant to give so much information that the reader will be able to distinguish between the different problems under which adapted indices must be used to compute agreement.

keywords: computer programs, nominal scale agreement, Cohen´s kappa.

authors adress:

Vakgroep Methoden & Technieken, Sociologisch Instituut, Rijks Universiteit Groningen, Oude Boteringestraat 23, 9712 GC Groningen.

1.0 PART I - AN OVERVIEW

1.1 Introduction

In this part theoretical information will be given on indices for computing agreement on nominal data, that will suffice for understanding the situations that are distinguished in part II, where computer programs will be compared. Here attention will be given to agreement indices for two raters, agreement indices for more than two raters, agreement indices for ratings, the comparison of lateral distributions, and the analysis on open-ended questions for two raters. The indices will not be discussed. This information cannot be considered as an introduction to nominal scale agreement. For such an introduction the reader is referred to Landis and Koch (1975), Bartko and Carpenter (1976), or Hollenbeck (1978).

1.2 Agreement indices for two raters

Still very often the proportion of observations, in which two raters agree in classifying the observations into the same category, is taken as an agreement index. This proportion is denoted as the "index of crude agreement" by Rogot and Goldberg (1966). The disadvantage of this index is that in case there are few categories the probability of equal assignments is greater than in case there are a lot of categories, especially when one takes into account the agreement that might be expected by chance. Bennett et al. (1954) have developed an "index of stability", S, in which there is a correction for the number of categories. They assume that all categories have equal chance to be used.

In reality however, it might be that all observations are assigned to only a few categories. This makes that S is not a satisfactory measure.

Therefore an "index of intercoder agreement", pi, was developed by Scott (1955). Here a correction is made for the number of categories and for the extent in which each category is used. Scott assumes that the distribution of the observations over the categories is known, and that it therefore applies for each category that the raters will both assign the same number of observations to a category. This might be a little different per rater, but in the long run, if the experiment is repeated enough times, it will be true.

This reasoning was objected by Cohen (1960). In his view the number of assignments per category per rater will be different most of the times, and merely be equal by incidence, "... one source of disagreement between a pair of judges is precisely their proclivity to distribute their judgements different over the categories" (p. 41). He has provided a measure in which this is taken into account. His

measure is called kappa, and is defined as "the proportion of agreement after chance agreement is removed from consideration" (p. 40), chance agreement here is based on the marginal distributions. At this moment Cohen´s kappa is one of the most frequently used agreement measures for nominal data. In a formula:

$$k = \{P(o) - P(e)\} / \{1 - P(e)\},$$

where P(o) is the observed proportion of agreement, and P(e) is the proportion of agreement expected by chance, based on the marginals.

Clement (1976) has provided a statistic which can be used in time series analysis. It takes into account the agreement on both occurrences and non-occurrences of the target behavior, while an adjustment is made for the frequencies of each. According to Clement the kappa statistic would "require the investigator to state in advance what the relative value of agreements for occurrences and non-occurrences would be, ..(Clement´s).. formula does not require such a priori judgments" (p. 258). The statistic is based on the assumption that the least frequent event will be weighted more heavily than the agreements for the most frequent event.

Several extensions of the kappa index have been made. Extensions to the situation where there are more than two raters are treated in the next sections.

Weights have been introduced by Cohen (1968). These are to be used when some notion of the seriousness of the rater´s disagreement is available. Cicchetti (1972) has introduced a linear weighting rationale to be used in case the data are rank ordered.

Another extension is that one can look at the agreement between the two raters, conditional on the classifications by one of these raters. It is possible to measure the agreement between both raters for only those observations that have been assigned by one of the raters to one specific category. This index was first given by Light (1971), based on an idea of Coleman.

Also several statistics, which have no connection with Cohen´s kappa, have been developed. One, proposed by Goodman and Kruskal (1954) is of the same type as the kappa coefficient. Their coefficient is based on optimal prediction. For each rater a marginal frequency is used that corresponds to a hypothesized modal class. By Light (1971) a statistic was provided for evaluating patterns of agreement. This statistic, A(p), is a function of the individual cell probabilities on the diagonal cells of the agreement table.

For the situation of binary data several measures have been developed. Dice (1945) has proposed a measure which estimates the conditional probability of agreement on the presence of an attribute of investigation, given the average of the proportion judged to be present by the two raters. For the situation in which the observation is judged to be

more often absent than present Dice proposed a measure which ignores the proportion in the cell on which both raters agree the observation is present. This measure is called Ap. For the reverse situation Dice proposed Aq.

A combined estimate of agreement based on the two estimates proposed by Dice, was presented by Rogot and Goldberg (1966). Their "index of adjusted agreement", A1, was developed independently from the one by Dice, in this index the expected amount of agreement is always 50 per cent. In addition Rogot and Goldberg also developed an "alternate index of agreement", A2, based on conditional probability.

Also available for binary data is the similarity ratio, or coefficient of Jaccard (Anderberg, 1973). In this index the ratio is expressed of observations in which the attribute of investigation is present according to both raters on observations in which the attribute is present according to at least one of the raters.

1.3 Agreement indices for more than two raters

When there are more than two raters to be compared, there are two definitions of agreement where to choose from.

One starts from the position that there is agreement if all raters agree in classifying an observation into the same category. This is called simultaneous agreement. Based on this definition Ross (1977) has developed a kappa statistic, he has also presented a weighted version of the statistic.

In the other situation the mean is taken of the kappa's computed between all pairs of raters. This is called pairwise agreement. A kappa index for this situation was provided by Light (1971). A better index was proposed by Hubert (1977a), who computed first the mean observed and mean expected agreement, and next entered these into the formula for kappa.

In the same article Hubert has presented a statistic in which one rater is considered as a standard and in which the other raters are compared pairwise with this standard. By Light (1971) a statistic, G, was provided for the comparison of a number of raters with a standard, which is not of the kappa type.

Weighted versions of the two statistics by Hubert can easily be provided. Only one matrix of weights has to be defined, which applies for all pairs of raters.

1.4 Agreement indices for ratings

For a discussion on the difference between the types of statistics for raters vs. ratings, see Conger (1980).

A kappa statistic for agreement in ratings was proposed by Fleiss (1971). Here the raters judging one subject are

not necessarily the same as those judging another subject. Also it is not necessary that each subject is rated the same number of times. In this statistic the expected chance agreement is expressed in another way than where a comparison of raters is concerned, it is based on the number of times a subject was assigned to a category.

In his article Fleiss has also presented how to measure the agreement within a particular category, this is intra-class agreement.

A weighted version of the statistics by Fleiss was presented by Schouten (1980, 1982), in these articles also an interclass kappa was presented, denoting the agreement between different categories.

For binary data Fleiss (1965) has proposed a nonparametric model for the errors underlying the judgments. He has given the conditions underr which Cochran's Q statistic is valid for testing the hypothesis of no systematic differences among the judgments of the different raters. Also for binary data Armitage et al. (1966) proposed indices of agreement. These are especially designed to be used in studies on recording of signs, this is e.g. the registration whether a specific kind of behavior happens in a certain time-interval or not. These indices are the mean "majority agreement index", in which the frequency is measured with which the raters agree with the majority opinion, the mean "pair disagreement index", measuring the frequency with which pairs of raters disagree, and the "standard deviation agreement index", measuring the variation between observations in the number of raters recording a positive finding.

Fleiss and Cuzick (1979) have considered the situation in which different observations are judged by different raters on a binary scale, and where the number of raters per observation varies.

1.5 Comparing lateral distributions

In case one wants to know something about the classifications on which raters do not agree, the distributions of the row and column totals of the agreement table, called lateral distributions, can be compared. For this purpose Maxwell (1970) has derived a test statistic based on estimation by the method of moments. His result enables the researcher to test whether the overall distribution of observations over categories between the two raters differ from each other.

By Fleiss and Everitt (1971) a method was provided to identify sources of significant differences between the two marginals of the agreement table. Here single categories can be investigated, but also a combination of categories, which will be considered as a new single category. Two major methods can be used for setting confidence limits on, or for testing hypotheses about y, where y is any linear combination of differences between corresponding marginals. The

multiple comparisons available for computing these confidence intervals are called the Scheffe type solution and the Bonferroni type solution.

Finally it is possible to investigate whether differences exist between corresponding non-diagonal cells. Here the null hypothesis is tested that the expected value, based on the marginals for the frequency in cell {i,j} of the agreement table equals the expected frequency for cell {j,i}, where i is unequal to j. Maxwell (1970, p. 653) states that "such a test would tell us whether the number of subjects about which the judges disagreed was distributed by them in a similar manner amongst the other available categories."

1.6 Analysis on open-ended questions for two raters

In this section measures are dealt with to be used in case categories have to be developed for a variable by the raters, and the observations must be assigned to these categories by these raters.

A rater can assign the observations belonging to a variable to categories of a scale. This might be a scale he developed himself. Another rater might have done the same task. Both raters may have used different scales, each also of a different number of categories. The assignments by both raters can be analyzed by means of a comparison of the classification of all pairs of observations. For each rater it holds that the two observations of a pair are classified into the same category, or into different ones. To give an example, suppose that four observations have been assigned as follows:

observation w	p1	q3
observation x	p1	q3
observation y	p2	q3
observation z	p2	q4

Both raters take the position that the observations w and x are assigned to the same category as separately developed. Concerning this pair there is complete agreement. Concerning the pair w and y this is not true. According to rater Q they must be assigned to the same category, but according to rater P this is not true. There is no agreement here. Finally it is possible that both raters take the position that the observations in a pair have to be assigned to different categories. In this case there is also agreement among the raters. This is true for the pair consisting of the observations w and z. Given the N observations, there are N(N-1)/2 pairs of observations. these can be placed in a table as given in Table 1.

Table 1 - Table resulting from the comparison of pairs of observations

		rater Q: same category	rater Q: different category
rater P	same category	a agreement (same)	b non-agreement
	different category	c non-agreement	d agreement (different)

N(N-1)/2

Agreement in cell a means agreement in classifying to the same categories, while agreement in cell d means agreement in classifying to different categories. For the data in Table 1 the meaning of the original categories is of no importance; this meaning may be different for both raters: not the verbal labeling but only the assignment result determines the outcome.

The statistics for agreement in the assignments can be distinguished to the criterium whether agreement is expressed in cell d of the table, or not. If so, three coefficients are available. The first one is a coefficient by Hubert (1977b), who called it gamma. The second coefficient, proposed by Montgommery and Crittenden (1977), is Yule's Q. Janson and Vegelius (1982) have critisized gamma, and have proposed their J-index as an agreement index.

Other coefficients are based on the relative frequency in cell a. The fraction in this cell is known as the dot-product. Several normalizations of the dot-product, and transformations of it to S1 and S2 types of scalability coefficients are discussed in Popping (1982).

2.0 PART II - COMPARISONS

2.1 Introduction

There are quite a lot of computer programs on nominal scale agreement. In this part the programs known to this author will be compared briefly. The comparisons will for a greater part be in terms of problems for which the programs can be used. The best known agreement measure by now is Cohen's kappa, as was noted in part I. Most programs are based on this measure. In case a program is based on an other measure than Cohen's kappa, this will be mentioned explicitely.

It is not the purpose to indicate which program is to be preferred. This may depend on the statistics the user needs, and on the computer that is available to him. Our comparison serves as an aid in quickly finding the program that is desired.

The programs are denoted by a name. In case the author had given a name to the program, this name is used. Otherwise a hypothetical name is given to the program, based on the name of the (first) author, and if this does not differentiate enough, the year of publication. The original names are followed by an asterix.

name	author(s)
AGREE3 *	Popping, 1981;
ANTON	Antonak, 1977;
BERK	Berk & Campbell, 1976;
CIC77	Cicchetti, Aivino & Vitale, 1977;
CIC78	Cicchetti, Lee, Fontana & Dowds, 1978;
CIC81	Cicchetti & Heavens, 1981;
CONGRU *	Watkins & McDermott, 1979;
CONTIN *	Vegelius, 1978;
DIAGNO *	Spitzer & Endicott, 1968;
GKAPPA *	Uebersax, 1981;
LARI	Larimer & Watkins, 1980;
MDC79B	McDermott & Watkins, 1979b;
MDC79C	McDermott & Watkins, 1979c;
RATCAT *	Cicchetti & Heavens, 1981;
STANDARD *	McDermott & Watkins, 1979a;
THOR	Thorton & Croskey, 1975;
WAT80	Watkins & Larimer, 1980;
WIX	Wixon, 1979.

In Heavens & Cicchetti (1978) and in Cicchetti & Heavens (1979) the same program is referred to. Therefore the first reference is not mentioned in the above list.

Nearly all information is taken from the program anouncements, not from the program descriptions. This implies that some information can be missing or misunderstood.

Except for the program DIAGNO all programs are general programs in which any datamatrix can be analyzed. DIAGNO assumes as input information concerning a standard scale for

 diagnosis of the American Psychiatric Association.

2.2 Comparison of problems that can be handled

In the programs several problems can be handled that refer to possibilities as dealt with in part I. The list which is mentioned below contains all problems for which computations can be performed. For the computations in all situations Cohen's kappa is used, unless it is mentioned that other indices are used. In 20 and 21 it is impossible to use kappa.

The problems are:

1. two raters are compared;
2. two raters are compared, weights are used;
3. two raters are compared, intraclass agreement is computed;
4. two raters are compared, intraclass agreement is computed, weighted;
5. two raters are compared, conditionalized on the categories of one of the raters;
6. two raters are compared, indices other than kappa are used;
7. more than two raters are compared, based on pairwise agreement;
8. more than two raters are compared, based on pairwise agreement, weighted;
9. more than two ratings are compared;
10. more than two ratings are compared, weighted;
11. more than two ratings are compared, indices other than kappa are used;
12. more than two ratings are compared, intraclass agreement;
13. more than two ratings are compared, interclass agreement;
14. a number of raters is pairwise compared with a standard;
15. a number of raters is pairwise compared with a standard, weighted;
16. a number of raters are compared with a standard, indices other than kappa are used;
17. simultaneous agreement is computed;
18. simultaneous agreement is computed, weighted;
19. computations on differences between pairs of kappas for two raters are performed;
20. lateral distributions are compared;
21. answers on open-ended questions are compared.

In the following tables the numbers in the columns refer to the problems mentioned above. An asterix means that the program mentioned in the row can handle the corresponding problem.

	1	2	3	4	5	6	7	8	9	10	11
AGREE3	*	*			*	*	*	*	*	*	*
ANTON	*					*					
BERK	*										
CIC77	*	*									
CIC78	*	*	*	*							
CIC81	*										
CONGRU									*		
CONTIN											
DIAGNO	*	*									
GKAPPA	*		*				*				
LARI	*					*					
MCD79B									*		
MCD79C											
RATCAT	*	*	*								
STANDARD											
THOR						*					
WAT80	*					*					
WIX	*										

	12	13	14	15	16	17	18	19	20	21
AGREE3	*	*	*	*		*	*		*	*
ANTON										
BERK										
CIC77										
CIC78										
CIC81								*		
CONGRU	*									
CONTIN										*
DIAGNO			*							
GKAPPA										
LARI										
MCD79B	*									
MCD79C					*					
RATCAT									*	
STANDARD					*					
THOR										
WAT80										
WIX										

Most programs can handle only few problems; the most extended is AGREE3, in this program only one problem is missing, that might be of importance, this is where intra-class agreement between raters is concerned.

In nearly all programs sampling characteristics, i.e., variance, etc., of the coefficients will be computed. These computations can not be performed in: CIC78, DIAGNO, and THOR. In GKAPPA the statistical significance of kappa is based on a Monte Carlo estimate.

2.3 Comparison on datasets

In some programs an agreement table must be entered, in others a datamatrix, i.e., a matrix with the raters in the column and the observations in the row. The fact that an agreement table is entered implies that only the classifications by two raters can be compared. There are also extended and mixed possibilities. Below they are listed:

1. one agreement table per run of the program;
2. several agreement tables after each other;
3. several agreement tables after each other, on each table several computations can be performed;
4. one datamatrix per run of the program;
5. one datamatrix for which all pairs of raters are compared;
6. several datamatrices after each other;
7. several datamatrices after each other, on each matrix several computations can be performed, also on parts of the datamatrix.

In the next table the numbers in the column refer to the possibilities mentioned above. An asterix means that the program mentioned in the row can handle the corresponding possibility.

	1	2	3	4	5	6	7
AGREE3			*				*
ANTON						*	
BERK		*					
CIC77					*		
CIC78					*		
CIC81				*			
CONGRU						*	
CONTIN				*		*	
DIAGNO				*			
GKAPPA					*		
LARI	*						
MCD79B				*			
MCD79C				*			
RATCAT	*						
STANDARD				*			
THOR						*	
WAT80							
WIX				*			

Most programs demand a datamatrix. AGREE3 can handle both agreement tables and datamatrices; it is possible to switch in one run of the program from one to the other.

The program RATCAT has the possibility to modify the input table.

The following programs are suitable for interactive use: AGREE3, GKAPPA, RATCAT, THOR, and WIX. AGREE3 can also be

used as a batch program, in which case the modifications of the data can not be done. If AGREE3 is used interactively the data can be modified in that part of the program where the classifications of answers on open-ended questions are compared.

2.4 On being friendly for users

Because we know all programs except AGREE3 only from the anouncements, it is very hard to say something about how friendly the programs are for the user. In case a mistake is made, AGREE3 will give a message. In case of a batch job, the program will then be terminated; in interactive use the user can continue there where the mistake was made.

In the input description of AGREE3 the formula's used in this program are described, further information on the (keyword) structure of the program is given (Popping, 1981). With regard to the other programs nothing is known about these subjects.

In the texts about the programs that have been used, nothing is noted concerning missing values in the datamatrices. AGREE3 can handle at most 40 different values in the range 1 - 999. Values not in this range are considered as missing, on occurrence listwise deletion is used. There are some possibilities to modify the data, see Popping (1981).

2.5 Comparison of program languages

Hereafter information is given about the languages in which the programs are written. Not only the language will be mentioned, but also the type of computer on which the programs are implemented, because there are several differences between versions of languages on the separate type of computers. This does not necessarily imply that programs cannot be converted from one system to another.

The program languages are:

1. fortran IV for CDC
2. fortran IV for IBM
3. fortran IV for HONEYWELL
4. fortran IV for UNIVAC
5. fortran V for CDC
6. fortran 10 for DEC
7. basic

In the next table the numbers in the column refer to the languages mentioned above. An asterix means that the program mentioned in the row is written in the language denoted in the column.

	1	2	3	4	5	6	7
AGREE3					*		
ANTON	(*)	(*)				*	
BERK	*	*		*			
CIC77		*					
CIC78		*					
CIC81		*					
CONGRU		*					
CONTIN		*					
DIAGNO		*					
GKAPPA	*						
LARI							*
MCD79B		*					
MCD79C		*					
RATCAT		*					
STANDARD		*					
THOR			*				
WAT80							*
WIX							*

Nearly all programs are written in fortran IV. In the table an asterix surrounded by brackets means that in the anouncements it was explitely mentioned that the program can be used on the corresponding type of computer.

The program WIX is written for a WANG computer, but can be used in basic.

2.6 Concluding remarks

It goes without saying that AGREE3 is the most extended program, which makes it attractive. A disadvantage is that the program needs a lot of central memory.

A following release of the program is planned for the winter of 1983, which version will be more appropriate for the common user, it will be programmed more efficiently, and will be extended with several new routines.

The program is available in LISTOR on the CDC 170/760 computer of the University of Groningen. It can be obtained from the LISTOR-group.

3.0 PART III - REFERENCES

3.1 References, part I

Anderberg, M.R. (1973) 'Cluster analysis for applications.' New York: Academic Press.

Armitage, P., Blendis, L.M. & Smyllie, H.C. (1966) 'The Measurement of observer disagreement in the recording of signs.' Journal of the Royal Statistical Society (A), 129, pp. 98 - 109.

Bartko, J.J. & Carpenter, W.T. (1976) 'Methods and theory of reliability.' Journal of Nervous and Mental Disease, 163, pp. 307 - 317.

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

Cicchetti, D.V. (1972) 'A new measure of agreement between rank ordered variables.' Proceedings of the 80th annual convention of the American Psychological Association, 7 pp. 17 - 18.

Clement, P.G. (1976) 'A formula for computing interobserver agreement.' Psychological Reports, 39, pp. 257 - 258.

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

Cohen, J. (1968) 'Weighted kappa. Nominal scale agreement with provision for scaled disagreement or partial credit.' Psychological Bulletin, 70, pp. 213 - 220.

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

Dice, L.R. (1945) 'Measures of the amount of ecological association between species.' Ecology, 26, pp. 297 - 302.

Fleiss, J.L. (1965) 'Estimating the accuracy of dichotomous judgments.' Psychometrika, 30, pp. 469 - 479.

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

Fleiss, J.L. & Cuzick, J. (1979) 'The reliability of dichotomous judgments: Unequal number of judges per subject.' Applied Psychological Measurement, 3, pp. 537 - 542.

Fleiss, J.L. & Everitt, B.S. (1971) 'Comparing the marginal

totals of square contingency tables.' British Journal of Mathematical and Statistical Psychology, 24, pp. 117 - 123.

Goodman, L.A. & W.H. Kruskal (1954) 'Measures of association for cross classifications.' Journal of the American Statistical Association, 49, pp. 732 - 764.

Hollenbeck, A.R. (1978) 'Problems of reliability in observational research.' In: Sacker, G.P. (ed.), Observing behavior, Volume 2. London: University Park Press, pp. 79 - 98.

Hubert, L.J. (1977a) 'Kappa revisited.' Psychological Bulletin, 84, pp. 289 - 297.

Hubert, L.J. (1977b) 'Nominal scale response agreement as a generalized correlation.' British Journal of Mathematical and Statistical Psychology, 30, pp. 98 - 103.

Janson, S. & Vegelius, J. (1982) 'The J-index as a measure of nominal scale response agreement.' Applied Psychological Measurement, 6, pp. 111 - 121.

Landis, J.R. & Koch, G.G. (1975) 'A review of statistical methods in the analysis of data arising from observer reliability studies.' Statistica Neerlandica, 29, pp. 101 - 123, 151 - 161.

Light, R.J. (1971) 'Measures of response agreement for qualitative data: Some generalizations and alternatives.' Psychological Bulletin, 76, pp. 365 - 377.

Maxwell, A.E. (1970) 'Comparing the classifications of subjects by two independent judges.' British Journal of Psychiatry, 116, pp. 651 - 655.

Montgommery, A.C. & Crittenden, K.S. (1977) 'Improving coding reliability for open-ended questions.' Public Opinion Quarterly, 41, pp. 235 - 243.

Popping, R. (1982) 'Traces of agreement. On the dot-product as a coefficient of agreement.' Quality and Quantity, 16, pp.

Rogot, E. & Goldberg, I.D. (1966) 'A proposed index for measuring agreement in test retest studies.' Journal of Chronic Diseases, 19, pp. 991 - 1006.

Ross, D.C. (1977) 'Testing patterned hypothesis in multi-way contingency tables using weighted kappa and weighted chi square.' Educational and Psychological Measurement, 37, pp. 291 - 307.

Schouten, H.J.A. (1980) 'Measuring pairwise agreement among many observers.' Biometrical Journal, 22, pp. 497 - 504.

Schouten, H.J.A. (1982) ´Measuring pairwise agreement among many observers. II. Some improvements and additions. Biometrical Journal, 24, pp. 431 - 435.

Scott, W.A. (1955) ´Reliability of content analysis: The case of nominal scale coding.´ Public Opinion Quarterly, 19, pp. 321 - 325.

3.2 References, part II

Antonak R.F. (1977) ´A computer program to compute measures of response agreement for nominal scale data obtained from two judges.´ Behavior Research Methods and Instrumentation, 9, p. 553.

Berk, R.A. & Campbell, K.L. (1976) ´A Fortran program for Cohen´s kappa coefficient of observer agreement.´ Behavior Research Methods and Instrumentation, 8, p. 396.

Cicchetti, D.V. (1977) Aivino, S.L. & Vitale, J. (1977) ´Computer programs for assessing rater agreement and rater bias for qualitative data.´ Educational and Psychological Measurement, 37, pp. 195 - 201.

Cicchetti, D.V. & Heavens, R. (1981) ´RATCAT (Rater agreement - Categorical Data).´ The American Statistician, 33, 1979, p. 91.

Cicchetti, D.V. & Heavens, R., ´A computer program for determining the significance of the difference between pairs of independently derived values of kappa or weighted kappa.´ Educational and Psychological Measurement, 41, pp. 189 - 193.

Cicchetti, D.V., Lee, C., Fontana, A.F. & Dowds, B.N. (1978) ´A computer program for assessing specific category agreement for qualitative data.´ Educational and Psychological Measurement, 38, pp. 805 - 813.

Heavens, R.H. (1978) Jr. & Cicchetti, D.V. (1978) ´A computer program for calculating rater agreement and bias statistics using contingency table input.´ Proceedings of the American Statistical Association, Statistical Computing Section, 21, pp. 366 - 370.

Larimer, L.D. & Watkins, M.W. (1980) ´A microcomputer basic program to calculate the level of agreement between two raters using nominal scale classification.´ Educational and Psychological Measurement, 40, pp. 773 - 775.

McDermott, P.A. & Watkins, M.W. (1979a) ´A Fortran program for testing agreement of multiple observers with a categorical standard on nominal scales.´ Educational and Psychologi-

cal Measurement, 39, pp. 669 - 672.

McDermott, P.A. & Watkins, M.W. (1979b) ´Program to evaluate general and conditional agreement among categorical assignments of many raters.´ Behavior Research Methods and Instrumentation, 11, pp. 399 - 400.

McDermott, P.A. & Watkins, M.W. (1979c) ´Computer program for assessing conjoint interrater agreement with a correct set of classifications.´ Behavior Research Methods and Instrumentation, 11, p. 607.

Popping, R. (1981) ´Computing agreement for qualitative data. AGREE 3.´ Bulletin no 44, Vakgroep Methoden en Technieken, Sociologisch Instituut, Groningen.

Spitzer, R.L. & Endicott, J. (1968) ´DIAGNO, A computer program for psychiatric diagnosis utilizing the different diagnostic procedure.´ Archives of General Psychiatry, 18, pp. 746 - 756.

Thorton, B.W. & Croskey, F.L. (1975) ´A computer program for calculating an index of interobserver reliability timeseries data.´ Educational and Psychological Measurement, 35, pp. 735 - 737.

Uebersax, J.S. (1981) ´GKAPPA. Generalized kappa coefficient.´ Applied Psychological Measurement, 5, p. 26.

Vegelius, J. (1978) ´CONTIN. A fortran IV program for nominal scale correlation coefficients.´ Educational and Psychological Measurement, 38, pp. 841 - 844.

Watkins, M.W. & Larimer, L.D. (1980) ´Interrater agreement statistics with the microcomputer.´ Behavior Methods and Research Instrumentation, 12, p. 466.

Watkins, M.W. & McDermott, P.A. (1979) ´A computer program for measuring levels of overall and partial congruence among multiple observers on nominal scales.´ Educational and Psychological Measurement, 39, pp. 235 - 239.

Wixon, D.R. (1979) ´Cohen kappa - coefficient of observer agreement, basic program for micro computers.´ Behavior Research Methods and Instrumentation, 11, p. 602.