Measures of agreement are especially of importance in the fields of biology, medicine and the social sciences. Most publications on this topic have appeared in biological and psychological journals. In 1975 a review has been published in which statistical methods are treated that can be used when analyzing data arising from observer reliability studies (LANDIS & KOCH, 1975a, 1975b). In this review a list of titles on measures of agreement is given. However since then nearly four years have passed and it seems worth while again paying attention to this topic: the G-index has been further developed, for interval data it is possible to base oneself on covariances, attention has been given to the problem of open ended questions.

In the bibliography that is presented here titles of articles and books are included in which the arising of measures of agreement is dealt with. Hardly any attention is given to texts in which these measures are used.

Some texts will be mentioned that do not deal with agreement itself, but that are closely related to it, or that are necessary to understand other texts on agreement.

Before presenting the titles some annotations will be made which should serve as a help in getting an impression of how this broad field is structured.

The titles do start from 1945 on. I do not pretend, however, that the bibliography is complete: 1) undoubtedly there are texts I do not know; 2) often it is hard to decide whether a text has to do with developments in the field of agreement, some decisions might be found dubious.

For data of a nominal level of measurement most attention is given today to measures of the kappa-type (COHEN, 1960; COHEN, 1968; FLEISS, 1971; LIGHT, 1971; FLEISS, 1975; LANDIS & KOCH, 1977a, 1977b; HUBERT, 1977). The last years a lot of attention is also

For data of an ordinal level of measurement Kendall's concordance test $W$ is used, and tests that are weighted versions of the kappa statistic (COHEN, 1968) and of the G-index (HOLLEY & KLINE, 1976; VEGELIUS, 1977a, 1977c).

To get a measure of agreement for data at an interval level of measurement the intraclass correlation coefficient, which is in fact a measure of association, is used. The researcher can use an analysis based on variance structures (EBEL, 1951), or an approach based on analysis of covariance (VAN DER KAMP & WELLENBERGH, 1976; WERTS et al., 1976).

Computer programs that can be used, are for data at the interval level the programs RELIABILITY in SPSS (SPRECHT & HOHLEN, 1977) and EBELREL, that has been developed at the Research Technische Dienstverlening, Subfaculteit Pedagogiek en Andragogiek of the Catholic University of Nijmegen. For ordinal data the program KENDW is available from the program library LISTOR at the University of Groningen. The one who has data at a nominal level can use the program COHEN (POPPING, 1977), that is also available in LISTOR. The author now works on an extended version of the COHEN-program and on a program that can be used when there are open ended questions, see BRENNAN & LIGHT, 1974 and MONTCOMBRY & CRITTENDEN, 1977.

general
ROBINSON, 1957; LIN, 1974; LANDIS & KOCH, 1975a, 1975b.

nominal data, general

nominal data, two judges

nominal data, more than two judges
CARTWRIGHT, 1956; FLEISS, 1965; ARMITAGE et al., 1966; BENNETT,
151

Of the G-index-type: HOLLEY & LIENERT, 1974.

nominal data, two judges, relative seriousness disagreements

nominal data, more than two judges, relative seriousness disagreements
Of the kappa-type: KLEIN et al., 1975; LIN, 1975; ROSS, 1977.
nominal data, conditional agreement
COLEMAN, 1966.

nominal data, comparison of one judge with a standard
WACKERLY et al., 1978.
nominal data, comparison of more than one judge with a standard
LIGHT, 1971.

nominal data, contribution of an extra judge
WILLIAMS, 1976.

nominal data, two judges, multiple diagnosis
FLEISS et al., 1972.

nominal data, multivariate agreement
KRIPPENDORFF, 1971; FELDMAN et al., 1972.

nominal data, number of ratings assessed
FLEISS, 1966; MAXWELL & PILLINER, 1968.

ordinal data
See also under nominal data, relative seriousness disagreements.

interval data, variance structures
EBEL, 1951; NYSTEDT, 1974.

interval data, covariance structures
VAN DER KAMP & WELLENBERGH, 1976; WERTS et al., 1976.

intraclass correlation

designing reliability studies
FLEISS, 1963; FLEISS et al., 1965; FLEISS, 1970.
ratings based on interview data

analyzing open ended questions
SHAPIRO, 1970; CRITTENDEN, 1971; BRENNAN & LIGHT, 1974;

computer programs
SPITZER & ENDICOTT, 1968; GREENE et al., 1975; THORTON & CROSKY,
1975; BERK & CAMPELL, 1976; CICCHETTI et al., 1976; POPPING,

bibliography

1. Ager, J.W., Jr. & Brent, S.B., An index of agreement between
a hypothesized partial order and an empirical rank order.

2. Alexander, W.H., The estimation of reliability when several
traits are available. Psychometrika, 12, 1947, pp. 79-99.

3. Armitage, P., Blendis, L.W. & Smyllie, H.C., The measurement
of observer disagreement in the recording of signs. J. Roy.

4. Arp, D.J., The problem of measurement and reliability with
special reference to the content analysis of psychothera¬
pic interviews. Ph. D., St. Louis, Missouri, 1968.


6. Bartko, J.J., The intraclass correlation coefficient as a
measure of reliability. Ps. Reports, 19, 1966, pp. 3-11.

7. Bartko, J.J., A note on the intraclass correlation coefficient
as a measure of reliability. Ps. Reports, 34, 1974, p. 418.

8. Bartko, J.J., On various intraclass correlation reliability

9. Bennett, B.M., Measures for clinicians' disagreement over signs.

10. Bennett, E.M., Blomquist, R.L. & Goldstein, A.C., Communicati¬
18, 1954, pp. 303-308.

Instr., 8, 1976, p. 396.

12. Brennan, R.L. & Light, R.J., Measuring agreement when to ob¬
servers classify people in categories not defined in advance.

13. Bruvold, W.H., Judgmental bias in the rating of attitude sta¬


51. Funkhouser, G.R. & Parker, E.B., Analyzing coding reliability-


70. Holley, J.W. & Sjöberg, L., Some characteristics of the G-index


84. Landis, J.R., A general methodology for the measurement of observer agreement when the data are categorical. Ph. D., University of North Carolina, 1975.


88. Landis, J.R. & Koch, G.G., An application of hierarchical kappa-type statistics in the assessment of majority agreement among


93. Light, R.J., Analysis of variance for catorial data with applications to agreement and association. Ph. D., Department of Statistics, Harvard University, 1969.


106. Popping, R., Cohen's kappa. Een coefficient voor interbeoor-


124. Swaminathan, H., Hambleton, R.K. & Algina, J., Reliability of criterion referenced tests: A decision-theoretic formula-


March 1979.