BIBLIOGRAPHY ON MEASURES OF AGREEMENT

roel popping

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 arizing 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 arizing 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

given to the G-index (HOLLEY & GUILFORD, 1964; LIENERT, 1973; VEGELIUS, 1976).

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 & MELLENBERGH, 1976; WERTS et al., 1976).

Computerprograms that can be used, are for data at the interval level the programs RELIABILITY in SPSS (SPECHT & 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 MONTGOMMERY & CRITTENDEN, 1977.

general

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

nominal data, general

BURDOCK et al., 1963; FLEISS, 1973; LANDIS, 1975.

nominal data, two judges

DICE, 1945; GOODMAN & KRUSKAL, 1954; BENNETT et al., 1954; SCOTT, 1955; ROGOT & GOLDBERG, 1966; RAE & TAYLOR, 1970.

Of the kappa-type: COHEN, 1960; EVERITT, 1968; FLEISS et al., 1969; KRIPPENDORFF, 1970; LIGHT, 1971; FLEISS, 1975; HUBERT, 1977.

Of the G-index-type: HOLLEY & GUILFORD, 1964; HOLLEY & SJOeBERG, 1968; LIENERT, 1973; VEGELIUS, 1976.

nominal data, more than two judges

CARTWRIGHT, 1956; FLEISS, 1965; ARMITAGE et al., 1966; BENNETT,

1972; EVERITT, 1977.

Of the kappa-type: FLEISS, 1971; LIGHT, 1971.

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

nominal data, two judges, relative seriousness disagreements

Of the kappa-type: COHEN, 1968; EVERITT, 1968; FLEISS et al., 1969; FLEISS & COHEN, 1973; HUBERT, 1978.

Of the G-index-type: HOLLEY & KLINE, 1976; VEGELIUS, 1977a, 1977c; VEGELIUS & JONSSON, 1977.

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.

Of the kappa-type: LIGHT, 1971; LIN, 1976.

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.

SIEGEL, 1956; CICCHETTI, 1972.

interval data, variance structures

EBEL, 1951; NYSTEDT, 1974.

interval data, covariance structures

VAN DER KAMP & MELLENBERGH, 1976; WERTS et al., 1976.

intraclass correlation

BARTKO, 1966; KRIPPENDORFF, 1970; FLEISS & COHEN, 1973; BARTKO, 1974, 1976.

designing reliability studies

FLEISS, 1963; FLEISS et al., 1965; FLEISS, 1970.

ratings based on interview data

FLEISS, 1970; CRITTENDEN & HILL, 1971.

analyzing open ended questions

SHAPIRO, 1970; CRITTENDEN, 1971; BRENNAN & LIGHT, 1974; MONTGOMMERY & CRITTENDEN, 1977.

computer programs

SPITZER & ENDICOTT, 1968; GREENE et al., 1975; THORTON & CROSKEY, 1975; BERK & CAMPBELL, 1976; CICCHETTI et al., 1976; POPPING, 1977; SPECHT & HOHLEN, 1977; CICCHETTI et al., 1978.

bibliography

1. Ager, J.W., Jr. & Brent, S.B., An index of agreement between a hypothesized partial order and an empirical rank order. J. Am. Stat. Ass., 73, 1978, pp. 827-830.
2. Alexander, W.H., The estimation of reliability when several traits are available. Psychometrika, 12, 1947, pp. 79-99.
3. Armitage, P., Blendis, L.M. & Smyllie, H.C., The measurement of observer disagreement in the recording of signs. J. Roy. Stat. Soc., A, 129, 1966, pp. 98-109.
4. Arp, D.J., The problem of measurement and reliability with special reference to the content analysis of psychotherapeutic interviews. Ph. D., St. Louis, Missouri, 1968.
5. Bahar, B.A., Recent research in reinterview procedures. J. Am. Stat. Ass., 63, 1968, pp. 41-63.
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 coefficients. Ps. Bul., 83, 1976, pp. 762-765.
9. Bennett, B.M., Measures for clinicans' disagreement over signs. Biometrics, 28, 1972, pp. 607-612.
10. Bennett, E.M., Blomquist, R.L. & Goldstein, A.C., Communications through limited response questioning. Publ. Op. Quart., 18, 1954, pp. 303-308.
11. Berk, R.A. & Campbell, K.L., A Fortran program for Cohen's kappa coefficient of observer agreement. Beh. Res. Meth. Instr., 8, 1976, p. 396.
12. Brennan, R.L. & Light, R.J., Measuring agreement when to observers classify people in categories not defined in advance. Br. J. Math. Stat. Ps., 27, 1974, pp. 154-163.
13. Bruvold, W.H., Judgmental bias in the rating of attitude statements. Ed. Ps. Meas., 35, 1975, pp. 605-611.

14. Burdock, E.I., Fleiss, J.L. & Hardesty, A.S., A new view of inter observer agreement. Personnel Ps., 16, 1963, pp. 373-384.

15. Cartwright, D.S., A rapid non-parametric estimate of multi-judge reliability. Psychometrika, 21, 1956, pp. 17-29.

16. Cicchetti, D.V., A new measure of agreement between rank ordered variables. Proceedings 80th Annual Convention, Am. Ps. Ass., 1972, pp. 17-18.

17. Cicchetti, D.V. & Allison, T., A new procedure for assessing reliability of scoring EEG sleep recordings. Am. J. EEG Techn., 11, 1971, pp. 101-109.

18. Cicchetti, D.V., Aivino, S.L. & Vitale, J., A computerprogram for assessing the reliability and systematic bias of individual measurement. Ed. Ps. Meas., 36, 1976, pp. 761-764.

19. Cicchetti, D.V., Lee, C., Fontana, A.F. & Dowdss, B.N., A computerprogram for assessing specific category rater agreement for qualitative data. Ed. Ps. Meas., 38, 1978, pp. 805-813.

20. Cochran, W.G., Errors of measurement in statistics. Technometrics, 10, 1968, pp. 637-666.

21. Cohen, J., A coefficient of agreement for nominal scales. Ed. Ps. Meas., 20, 1960, pp. 37-46.

22. Cohen, J., Weighted kappa. Nominal scale agreement with provision for scaled disagreement or partial credit. Ps. Bul., 70, 1968, pp. 213-220.

23. Cohen, J., R(c): a profile similarity coefficient invariant over variable reflection. Ps. Bul., 71, 1969, pp. 281-284.

24. Cohen, J., Weighted chi square: an extension of the kappa method. Ed. Ps. Meas., 32, 1972, pp. 61-74.

25. Coleman, J.S., Measures of concordance between members of social groups. Unpublished manuscript. Johns Hopkins University, 1966.

26. Crittenden, K.S., Actual and reconstructed coding procedure. In: McGee (ed.), Academic Janus. San Francisco, 1971.

27. Crittenden, K.S. & Hill, R.J., Coding reliability and validity of interview data. Am. Soc. Rev., 36, 1971, pp. 1073-1080.

28. Cronholm, J.N., A pair of non-parametric indices of agreement and disagreement. Report nr. 592, Ps. Div. U.S. Army, Med. Research Lab., Fort Knox, Kentucky, 1963.

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

30. Ebel, R.L., Estimation of the reliability of ratings. Psychometrika, 16, 1951, pp. 407-424.

31. Einhorn, H.J., Hogarth, R.M. & Klempner, E., Quality of group judgment. Ps. Bul., 84, 1977, pp. 158-172.

32. Everitt, B.S., Moments of the statistics kappa and weighted kappa. Br. J. Math. Stat. Ps., 21, 1968, pp. 97-103.

33. Everitt, B.S., Some properties of statistics used for measuring observer agreement in the recording of signs. Br. J. Math. Stat. Ps., 30, 1977, pp. 227-233.

34. Feldman, S., Klein, D.F. & Honigfeld, G., The reliability of a decision tree technique applied to psychiatric diagnosis. Biometrics, 28, 1972, pp. 831-840.

35. Finn, R.H., A note on estimating the reliability of categorial data. Ed. Ps. Meas., 30, 1970, pp. 71-76.

36. Flanders, N.A., The problem of observer training and reliability. In: Amidon & Hough (eds.), Interaction analysis: theory, research, and application. Reading, 1967.

37. Fleiss, J.L., Determination of the reliability of ratings by means of incomplete block designs. Am. Ps., 18, 1963, p. 420.

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

39. Fleiss, J.L., Assessing the accuracy of multivariate observations. J. Am. Stat. Ass., 61, 1966, pp. 403-412.

40. Fleiss, J.L., Estimating the reliability of interview data. Psychometrika, 35, 1970, pp. 143-162.

41. Fleiss, J.L., Measuring nominal scale agreement among many raters. Ps. Bul., 76, 1971, pp. 378-382.

42. Fleiss, J.L., Statistical methods for rates and proportions. New York, 1973.

43. Fleiss, J.L., Measuring agreement between two judges on the presence or absence of a trait. Biometrics, 31, 1975, pp. 651-659.

44. Fleiss, J.L. & Cohen, J., The equivalence of weighted kappa and the intraclass correlation coefficient as a measure of reliability. Ed. Ps. Meas., 33, 1973, pp. 613-619.

45. Fleiss, J.L., Cohen, J. & Everitt, B.S., Large sample standard errors of kappa and weighted kappa. Ps. Bul., 72, 1969, pp. 323-327.

46. Fleiss, J.L. & Everitt, B.S., Comparing the marginal totals of chi square contingency tables. Br. J. Math. Stat. Ps., 24, 1971, pp. 117-123.

47. Fleiss, J.L., Spitzer, R.L. & Burdock, E.I., Estimating accuracy of judgment using recorded interviews. Arch. Gen. Psychiatry, 12, 1965, pp. 562-567.

48. Fleiss, J.L., Spitzer, R.L., Endicott, J. & Cohen, J., Quantification of agreement in multiple psychiatric diagnosis. Arch. Gen. Psychiatry, 26, 1972, pp. 168-171.

49. Forge, R. la, Components of reliability. Psychometrika, 30, 1965, pp. 187-195.

50. Freeman, J. & Butler, E.W., Some sources of interview variance in surveys. Publ. Op. Quart., 40, 1976, pp. 79-91.

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

ty: the random-systematic-error coefficient. Publ. Op. Quart., 32, 1968, pp. 122-128.

52. Gallhofer, I.N., Coders' reliability in the study of decision-making concepts: replication in time and across topics. Methoden en Data Nieuwsbrief, 3, 1978, 1, pp. 58-74.

53. Garner, W.R. & Hake, H.W., The amount of information in absolute judgments. Ps. Review, 58, 1951, pp. 446-459.

54. Gleser, G.C., Cronbach, L.J. & Rajaratnam, N., Generalizing of scores influenced by multiple sources of variance. Psychometrika, 30, 1965, pp. 395-418.

55. Goodman, L.A. & Kruskal, W.H., Measures of association from cross classifications. J. Am. Stat. Ass., 49, 1954, pp. 732-764.

56. Greene, J.F., McCook, W.M. & Archambault, A.X., A computer-program to calculate adjusted and unadjusted interrater reliabilities for sets and subsets of judges. Ed. Ps. Meas., 35, 1975, pp. 689-691.

57. Grubs, F.E., Errors of measurement, precision, accuracy and the statistical comparison of measuring instruments. Technometrics, 15, 1973, pp. 53-66.

58. Guetzkow, H., Uniting and categorizing problems in coding qualitative data. J. Clin. Ps., 6, 1950, pp. 47-58.

59. Gulliksen, H., Methods for determining equivalence of measures. Ps. Bul., 70, 1968, pp. 534-544.

60. Guttman, L., A basis for analyzing test-retest reliability. Psychometrika, 10, 1945, pp. 255-282.

61. Guttman, L., The test-retest reliability of qualitative data. Psychometrika, 11, 1946, pp. 81-95.

62. Haggard, E.A., Intraclass correlation and the analysis of variance. New York, 1958.

63. Hollander, M. & Sethuraman, J., Testing for agreement between two groups of judges. Biometrika, 65, 1978, pp. 403-411.

64. Holley, J.W., A note on the relationship between R and Q factors. Scan. J. Ps., 5, 1964, pp. 143-149.

65. Holley, J.W., A reply to Philip Levy. Scan. J. Ps., 7, 1966, pp. 313-317.

66. Holley, J.W., Philip Levy and the G-index. A final reply. Scan. J. Ps., 8, 1967, p. 250.

67. Holley, J.W. & Guilford, J.P., A note on the G-index of agreement. Ed. Ps. Meas., 24, 1964, pp. 749-753.

68. Holley, J.W. & Kline, P., On the generalization of the G-index of agreement: G(O) for use with ordinal scores. Scan. J. Ps., 17, 1976, pp. 149-152.

69. Holley, J.W. & Lienert, G.A., The G-index of agreement in multiple ratings. Ed. Ps. Meas., 34, 1974, pp. 817-822.

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

of agreement. Mult. Beh. Research, 1, 1968, pp. 107-114.

71. Horst, P.A., A generalized expression for the reliability of measures. Psychometrika, 14, 1949, pp. 21-31.

72. Hubert, L.J., A relationship between the assignment problem and some simple statistical techniques. Quality and Quantity, 10, 1976, pp. 341-348.

73. Hubert, L.J., Kappa revisited. Ps. Bul., 84, 1977, pp. 289-305.

74. Hubert, L.J., A general formula for the variance of Cohen's weighted kappa. Ps. Bul., 85, 1978, pp. 183-184.

75. Huynh, H., Reliability of decisions in domain-referenced testing. J. Ed. Meas., 13, 1976, pp. 253-264.

76. Huynh, H., Reliability of multiple classifications. Psychometrika, 43, 1978, pp. 317-325.

77. Kamp, L.J.Th. van der, Studies in reliability. Leiden, 1974.

78. Kamp, L.J.Th. van der & Mellenbergh, G.J., Agreement between raters. Ed. Ps. Meas., 36, 1976, pp. 311-317.

79. Klein, D.F., Ross, D.C. & Feldman, S., Analysis and display of psychopharmalogical data. J. Psychiatr. Research, 12, 1975, pp. 125-147.

80. Koch, G.G., Landis, J.R., Freeman, J.L., Freeman, H.D., Jr. & Lehnen, R.G., A general methodology for the analysis of experiments with repeated measurement of categorial data. Biometrics, 33, 1977, pp. 133-158.

81. Krippendorff, K., Bivariate agreement coefficients for reliability of data. In: Borgatta & Bohrnstedt (eds.), Sociological methodology 1970. San Francisco, 1970.

82. Krippendorff, K., Estimating the reliability, systematic error and random error of interval data. Ed. Ps. Meas., 30, 1970, pp. 61-70.

83. Krippendorff, K., Reliability of recording instructions: multivariate agreement for nominal data. Behavioral Science, 16, 1971, pp. 228-235.

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

85. Landis, J.R. & Koch, G.G., A review of statistical methods in the analysis of data arising from observer reliability studies (part I). Stat. Neerl., 29, 1975, pp. 101-123, a.

86. Landis, J.R. & Koch, G.G., A review of statistical methods in the analysis of data arising from observer releability studies (part II). Stat. Neerl., 29, 1975, pp. 151-161, b.

87. Landis, J.R. & Koch, G.G., The measurement of observer agreement for categorial data. Biometrics, 33, 1977, pp. 159-174, a.

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

multiple observers. Biometrics, 33, 1977, pp. 363-374, b.

89. Landis, J.R., Stanish, W.H., Freeman, J.L. & Koch, G.G., A computerprogram for the generalized chi square analysis of categorial data using weighted least squares (GENCAT). Computer Programs in Biomedicine, 6, 1976, pp. 196-231.

90. Levy, P., Properties of the Holley-Guilford G-index of agreement in R and Q factor analysis. Scan. J. Ps., 7, 1966, pp. 239-243.

91. Levy, P., A reply to Jasper Holley's defense of the G-index. Scan. J. Ps., 8, 1967, p. 38.

92. Lienert, G.A., On a generalized G-index of agreement. Ed. Ps. Meas., 33, 1973, pp. 767-772.

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

94. Light, R.J., Measures of response agreement for qualitative data: some generalizations and alternatives. Ps. Bul., 76, 1971, pp. 365-377.

95. Lin, Y.S., Statistical measurement of agreement. Ph. D., School of Statistics, University of Minesota, 1974.

96. Lin, Y.S., Weighted kappa for three raters. Am. Stat. Ass., Proceedings Soc. Science Section, 1975, pp. 529-531.

97. Lin, Y.S., Conditional agreement score with two raters. Am. Stat. Ass., Proceedings Soc. Science Section, 1976, pp. 548-550.

98. Loevinger, J., A systematic approach to the construction and evaluation of tests of ability. Ps. Monogr., 61, 1947, no. 4.

99. Loevinger, J., The technique of homogenous tests compared with some aspects of 'scale analysis' and factor analysis. Ps. Bul., 45, 1948, pp. 507-530.

100. Maxwell, A.E., Comparing the classifications of subjects by two independent judges. Br. J. Psychiatry, 116, 1970, pp. 651-655.

101. Maxwell, A.E. & Pilliner, A.E.G., Deriving coefficients of reliability and agreement for ratings. Br. J. Math. Stat. Ps., 21, 1968, pp. 105-116.

102. Mellenbergh, G.J., Een onderzoek naar het beoordelen van open vragen. Ned. Tijdschr. Ps., 26, 1971, pp. 102-120.

103. Montgommery, A.C. & Crittenden, K.S., Improving coding reliability for open ended questions. Publ. Op. Quart., 41, 1977, pp. 235-243.

104. Nystedt, L., Consensus among judges as a function of amount of information. Ed. Ps. Meas., 34, 1974, pp. 91-101.

105. Overall, J.E., Estimating individual rater reliabilities from analysis of treatment effects. Ed. Ps. Meas., 28, 1968, pp. 255-264.

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

delaarsbetrouwbaarheid voor nominale data. Bulletin no. 19, Vakgroep Methoden & Technieken, Sociologisch Instituut. Groningen, july, 1977.

107. Popping, R., Comment on an article on measures of response agreement by Light. Bulletin no 29, Vakgroep Methoden & Technieken, Sociologisch Instituut, Groningen, januari 1979.

108. Rae, D.W. & Taylor, M., The analysis of political cleavages. New Haven, 1970, pp. 115-145.

109. Robinson, W.S., The statistical measurement of agreement. Am. Soc. Rev., 22, 1957, pp. 17-25.

110. Rogot, E. & Goldberg, I.D., A proposed index for measuring agreement in test-retest studies. J. Chronic Diseases, 19, 1966, pp. 991-1006.

111. Ross, D.C., Testing patterned hypothesis in multi-way contingency tables using weighted kappa and weighted chi square. Ed. Ps. Meas., 37, 1977, pp. 291-307.

112. Schutz, W.C., Reliability, ambiguity and content analysis. Ps. Review, 59, 1952, pp. 119-129.

113. Schutz, W.C., On categorizing qualitative data in content analysis. Publ. Op. Quart., 22, 1958, pp. 503-515.

114. Schott, W.A., Reliability of content analysis: the case of nominal scale coding. Publ. Op. Quart., 19, 1955, pp. 321-325.

115. Shapiro, M.J., Discovering interviewer bias in open ended survey responses. Publ. Op. Quart., 34, 1970, pp. 412-415.

116. Siegel, S., Nonparametric statistics for the behavioral sciences. New York, 1956, pp. 229-239.

117. Sjöberg, L. & Holley, J.W., A measure of similarity between individuals when scoring directions are arbitrary. Mult. Beh. Research, 2, 1967, pp. 377-384.

118. Smith, D.E., The social construction of documentary reliability. Social Inquiry, 44, 1974, pp. 257-268.

119. Specht, D.A. & Hohlen, M., SPSS Reliability: subprogram for item and scale analysis. Evanston, 1975.

120. Spiegelman, M., Terwilliger, C. & Fearing, F., The reliability of agreement in content analysis. J. Soc. Ps., 37, 1953, pp. 175-187.

121. Spitzer, R.L., Cohen, J., Fleiss, J.L. & Endicott, J., Quantification of agreement in psychiatric diagnosis. A new approach. Arch. Gen. Psychiatry, 17, 1967, pp. 83-87.

122. Spitzer, R.L. & Endicott, J., Diagno: A computer program for psychiatric diagnosis utilizing the differential diagnostic procedure. Arch. Gen. Psychiatry, 18, 1968, pp. 746-756.

123. Summers, J.O. & McKay, D.B., On the validity and reliability of direct similarity judgments. J. Marketing Research, 13, 1976, pp. 289-295.

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

tion. J. Ed. Meas., 11, 1974, pp. 263-267.

125. Thorton, B.W. & Croskey, F.L., A computer program for calculating an index of interobserver reliability from timeseries data. Ed. Ps. Meas., 35, 1975, pp. 735-737.

126. Vegelius, J., On the generalization of the G-index. Ed. Ps. Meas., 36, 1976, pp. 595-600.

127. Vegelius, J., An ordinal scale generalization of the G-index invariant over item reflection. Ed. Ps. Meas., 37, 1977, pp. 31-35, a.

128. Vegelius, J., The problem of varying scales for G-index generalization. Ed. Ps. Meas., 37, 1977, pp. 279-282, b.

129. Vegelius, J., On the weighted G-index. Ed. Ps. Meas., 37, 1977, pp. 839-842, c.

130. Vegelius, J., On the utility of the E-correlation coefficient concept in psychological research. Ed. Ps. Meas., 38, 1978, pp. 605-611, a.

131. Vegelius, J., Contin. A Fortran IV program for nominal scale correlation coefficients. Ed. Ps. Meas., 38, 1978, pp. 841-844, b.

132. Vegelius, J. & Jonsson, H., On weighted G analysis. Mult. Beh. Research, 12, 1977, pp. 243-254.

133. Wackerly, D.D., McClave, J.T. & Rao, P.V., Measuring nominal scale agreement between a judge and a known standard. Psychometrika, 43, 1978, pp. 213-224.

134. Walsh, J.E., Concerning the effect of intraclass correlation on certain significance tests. Ann. Math. Stat., 18, 1947, pp. 88-96.

135. Werts, C.E., Jöreskog, K.G. & Linn, R.L., Analyzing ratings with correlated intrajudge measurement errors. Ed. Ps. Meas., 36, 1976, pp. 319-328.

136. Williams, G.W., Comparing the joint agreement of several raters with another rater. Biometrics, 32, 1976, pp. 619-627.

137. Wing, J.K., Birley, J.L.T., Cooper, J.E., Graham, P. & Isaacs, A.D., Reliability of a procedure for measuring and classifying 'present psychiatric state'. Br. J. Psychiatry, 113, 1967, pp. 499-515.

138. Woodward, J.L. & Franzen, R., A study on coding reliability. Publ. Op. Quart., 12, 1948, pp. 253-257.

March 1979.