KM 19(1985)
pag 105-114



# A log-linear approach to the inter-judge reliability of qualitative judgments

Willem A. van der Kloot

Tineke M. Willemsen

Abstract

In a typical observation study, a number of judges categorize a large number of behavioral acts of the observed persons. When one is interested in how often an act occurs in a particular category, one may define inter-judge reliability as the agreement between the judges' distributions of acts over the categories. If there are many judges and many observation conditions, log-linear analysis can be used to fit models that involve judges interactions and models that do not involve such interactions. If the latter models fit equally well, the judgments can be regarded as reliable. The indices of normed fit and of non-normed fit (cf. Bonett and Bentler, 1983) can be used to compare models.

Vakgroep Methoden en Technieken, Subfaculteit Psychologie,
Rijksuniversiteit Leiden, Hooigracht 15, 2312 KM Leiden,
telefoon 071-148333, toestel 5130 of 6354.

Reliability (also: dependability, consistency, stability; cf. Guilford, 1954) is usually operationalized in terms of repeatability, i.e. repeatability over time (test-retest reliability), repeatability over instruments (parallel test reliability) and repeatability over observers (inter-judge reliability). The latter operationalization is used when the data consist of judgments made by different individuals (judges, observers). Such data are said to be more reliable when there is less variation between the judges. As long as the judgments are quantitative or numerical (as in grade points or ratings on rating scales) various univariate and multivariate methods can be used to study the reliability of those judgments.

The choice of a particular method depends of course on the type of judgments and on the circumstances under which they are collected. Research with rating scales usually yields data that are 'three-way/three-mode' in that they involve a number of **judges** who judge a number of **stimuli** (e.g. other persons, objects) on a number of **variables** (attributes, scales). Some studies involve an even larger number of facets as the judgments are collected under different tasks, instructions, or other conditions. In such cases one can decompose the total variation into several components. Some of these components represent systematic or wanted variation, i.e. the kind of differences that one wishes to be large. Other components represent unwanted or error variation; they reflect those differences that should be minimal if the judgments are to be reliable. Depending on which particular components are regarded as systematic variation and which other sources are regarded as error variation, it is possible to estimate several coefficients of reliability, or rather, generalizability. This approach is used in the so-called generalizability theory (Gleser, Cronbach, and Rajaratnam, 1965; Cronbach, Gleser, Nanda, and Rajaratnam, 1972).

If the observations consist of qualitative judgments (e.g. categorizations of behavioral acts) inter-judge reliability is harder to assess, especially if there are more than two judges and if the judgment procedure is repeated (e.g. under different conditions). The simplest case is a study in which two judges classify a number of observations into a set of mutually exclusive categories. The data can then be arranged in a categories-of-judge-one × categories-of-judge-two frequency table, and the inter-judge reliability can then be expressed by some measure of association: for instance a Phi-coefficient or tetrachoric correlation when the data are dichotomous or Kappa (Cohen, 1960, 1968) when the judgments are polytomous

(for reviews of inter-judge association measures see Fleiss, 1975; Landis and Koch, 1975a, 1975b).

When there are more than two judges and more observation conditions it is possible to compute association measures for every pair of judgments (i.e. judgments given by two different judges or judgments collected at two different occasions). An average of these pairwise associations is then usually taken as 'the' reliability of the observations. This procedure has at least three disadvantages:

1. It is possible to compute many reliability coefficients since there are many ways in which one may average the pairwise coefficients.

2. More or less complex interactions between judges and observation conditions may remain undetected.

3. Association measures as Phi or Kappa express the act-by-act correspondence of the pairs of judgments. This is sometimes a much too severe criterion for assessing reliability as one is often not interested in the reliability of single observations.

In a typical observation study each judge classifies a relative large number of acts of an observed person. The final result is an array of frequencies, for instance: judge J classified the behavior of boy P 12 times as kicking, 7 times as laughing, 5 times as yelling, etc. If one is only interested in how often a certain behavior occurs, and not in sequences of acts, it would be sufficient for two judges to correspond with respect to the numbers of acts they place in the same observation category. In such cases it is not necessary that they classify each individual act in the same manner. An appropriate index of inter-judge reliability can then be based on the similarity between the category frequencies of the two judges. A possible measure is the value of the $\chi^2$-test for independence. If the $\chi^2$-value is high it means that that there is an **interaction** between the judges and the categories in that the judges' distributions of acts over categories are different. If the $\chi^2$-value is low, such an interaction is less likely.

The above approach can be easily extended to cases where there are many judges and many observation conditions. The judgments can then be arranged in a judges $\times$ categories $\times$ conditions frequency table that can be analyzed by means of standard log-linear methods (cf. Bishop, Fienberg, and Holland, 1975; Everitt, 1977). Such methods estimate parameters for the marginal distributions of the variables (i.e. judges, categories, conditions) and for the first- and higher-order interactions among them. There are two sorts of interactions: those that involve the judges and those that do not involve them. As interactions involving the judges reflect inter-judge differences,

reliability can be expressed in terms of the relative importance of those interactions. A general procedure would be to compare the (lack of) fit of a log-linear model that includes judge interaction parameters with the (lack of) fit of the same model without those parameters. If the latter model fits about equally well, then the observations can be regarded as 'reliable'. If the first model has a considerably better fit, then the judgments are to some extent 'unreliable'.

A more detailed exposition of this approach will be given in the next sections of this paper, in which we describe an application of log-linear methods to the categorizations of verbal acts.

## Data

The data for the present study were obtained from four judges who each judged the 1565 scorable acts that were exchanged among the members of a five person discussion group[1]). These acts were judged using Bales' (1950, 1970) system of interaction process analysis (IPA) in which each act is classified according to its initiator, its target and its content. The initiator variable has five categories denoting the five group members. The target variable has six categories, i.e. the five group members and the group as a whole. The content categories were the original 12 IPA categories. For the present analysis they were lumped together in the four areas distinguished by Bales: (a) social emotional positive (IPA: 1, 2, 3), (b) task-oriented: answers (IPA: 4, 5, 6), (c) task-oriented: questions (IPA: 7, 8, 9), and (d) social emotional negative (IPA: 10, 11, 12). The group discussion was subdivided in five periods of about equal duration, containing respectively 291, 356, 281, 321, and 316 acts. These subsets of acts were judged on five occasions each approximately two weeks apart.

The judgments were arranged in a $5 \times 4 \times 5 \times 6 \times 4$ Periods $\times$ Judges $\times$ Initiator $\times$ Target $\times$ Category table. We call this the PJITC table. Cell ijklm of this table contains the number of acts in period i classified by judge j as directed from group member k to group member l in content category m. Note that the periods and judges variables are fixed in the sense that the number of acts judged by each judge in each period is determined by the design of the study.

[1] See for a detailed description of the group discussion and the judging procedure Willemsen (1984).

## Results

From the complete PJITC table three other tables were derived by ignoring either I, T, or C. This was done in order to avoid small cell entries in the tables to be analyzed (the complete table contains 2400 cells, and we have 'only' 6260 observations). The resulting tables will be called PJTC, PJIC, and PJIT. These tables were analyzed by means of the log-linear analysis program of BMDP (Brown, 1981). For each table three series of analyses were performed: one on the complete table, and two analyses on tables derived by randomly dividing the 1565 acts into two approximately equal subsets. This was done in order to study the stability of the best-fitting models.

The models considered here varied from a completely saturated model (for instance: {PJTC}) which contains parameters for all zero- and higher-order interactions, to a maximally constrained model. The maximally restricted model is in all cases {PJ}. This model contains only the parameters that follow necessarily from fixing P and J and their interaction[2]). In between the least and most restricted models are models of varying complexity. These intermediate models fall into two classes: models that contain free parameters for the interaction between J and one or more of the other variables, and models that do not contain such parameters. The relative fit of the latter models indicates the (un)importance of the parameters for interactions involving J: the higher this relative fit, the higher the inter-judge reliability. A measure of relative fit is obtained by comparing the fit of the completely saturated model with the fit of the same model after it is restricted by excluding one or more of the free parameters involving J. The most restricted models in this regard are {PJ,PTC} for the PJTC table, and {PJ,PIC} and {PJ,PIT} for the other tables. These models imply that the distribution of acts over the combinations of P, T, and C (resp. P, I, and C, and P, I, and T) is similar for each judge. If such models are tenable (i.e. not less tenable than models that do include interactions with J) it would imply that the inter-judge differences are not so big that they would lead to wrong conclusions regarding the other effects present in the data. In this sense the judgments are then reliable.

[2] In this paper we use the standard notation of hierarchical log-linear models. {ABC} stands for a model that contains parameters for A, B, C, AB, AC, BC, and ABC. The inclusion of an interaction term in a model means that the parameters for all lower-order interactions are included in the model. For example the model {PJ, TC} contains P, J, T, C, PJ, and TC.

Table 1, 2, and 3 display the results of our analyses. The intermediate models consist of the restricted models (i.e. {PJ,PTC}, {PJ,PIC} and {PJ,PIT}) plus those judge interactions that were found to be significant in a number of exploratory analyses in which the separate contributions of all variables and interactions were estimated. The significant interactions were JT, JC and PJT. estimated. In Tables 1, 2, and 3, next to each model are printed the degrees of freedom, the likelihood ratio statistic G-square, and two indices $\hat{\Delta}$ and $\hat{\delta}$. These latter statistics are proposed by Bonett and Bentler (1983) as ways of gauging the importance of the parameters in a model. $\Delta$ is called the index of normed fit; it measures the relative loss reduction obtained by adding parameters to the most restricted model. $\delta$ is the index of non-normed fit; the difference with the former measure is that in $\delta$ the losses are divided by their respective degrees of freedom. For instance, for model {PJ,PIT} (see Table 3) these values are computed in the following manner:

$$\hat{\Delta}(PJ,PIT) = \frac{G^2(PJ) - G^2(PJ,PIT)}{G^2(PJ)} = \frac{5296.6 - 300.0}{5296.6} = .943$$

$$\hat{\delta}(PJ,PIT) = \frac{\{G^2(PJ)/df(PJ)\} - \{G^2(PJ,PIT)/df(PJ,PIT)\}}{G^2(PJ)/df(PJ)} =$$

$$= \frac{(5296.6/580) - (300.0/435)}{5296.6/580} = .925$$

These two values can be regarded as reliability measures for the PJIT data, although they are not the same as reliability coefficients in classical test theory. As Bonett and Bentler state:

> Although $\Delta$ is bounded by zero and unity, the normed fit index is not a correlation coefficient and does not provide a "proportion of variance accounted for" interpretation. The value of $\Delta$ simply reflects the percent improvement in formal goodness of fit of Mb [the model under scrutiny] over Mn [the most restricted model] (p. 157).

With regard to $\delta$ they note that

> . . . it has an upper bound of 1.0 but will not necessarily be positive . . . . In contrast to the normed fit index, $\delta$ can decrease in value when restrictions are lifted from the model if the improvement in goodness of fit is not commensurate with the loss of degrees of freedom. This property becomes very useful in exploratory model selection. ... a given parameter should not be added if $\delta$ decreases when that parameter is added (p.158).

TABLE 1

Results of the Log-linear Analysis of the

Periods × Judges × Initiators × Categories Table

| Model | df | full table | | | sample 1 | | | sample 2 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $G^2$ | $\hat{\Delta}$ | $\hat{\delta}$ | $G^2$ | $\hat{\Delta}$ | $\hat{\delta}$ | $G^2$ | $\hat{\Delta}$ | $\hat{\delta}$ |
| PJ | 380 | 3591 | | | 1825 | | | 1990 | | |
| PJ,PIC | 285 | 165 | .954 | .939 | 114 | .938 | .917 | 125 | .937 | .916 |
| PJ,JC,PIC | 276 | 132 | .963 | .950 | 99 | .946 | .926 | 105 | .947 | .927 |

TABLE 2

Results of the Log-linear Analysis of the

Periods × Judges × Targets × Categories Table

| Model | df | full table | | | sample 1 | | | sample 2 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $G^2$ | $\hat{\Delta}$ | $\hat{\delta}$ | $G^2$ | $\hat{\Delta}$ | $\hat{\delta}$ | $G^2$ | $\hat{\Delta}$ | $\hat{\delta}$ |
| PJ | 460 | 5446 | | | 2623 | | | 3030 | | |
| PJ,PTC | 345 | 443 | .919 | .892 | 274 | .896 | .861 | 287 | .905 | .874 |
| PJ,JT,PTC | 330 | 386 | .929 | .901 | 219 | .916 | .883 | 264 | .913 | .879 |
| PJ,JC,JT,PTC | 321 | 347 | .936 | .909 | 203 | .923 | .889 | 242 | .920 | .885 |
| PJ,PJT,PTC | 270 | 280 | .949 | .913 | 191 | .927 | .876 | | | |

TABLE 3

Results of the Log-linear Analysis of the

Periods × Judges × Initiators × Targets Table

| Model | df | full table | | | sample 1 | | | sample 2 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $G^2$ | $\hat{\Delta}$ | $\hat{\delta}$ | $G^2$ | $\hat{\Delta}$ | $\hat{\delta}$ | $G^2$ | $\hat{\Delta}$ | $\hat{\delta}$ |
| PJ | 580 | 5297 | | | 2511 | | | 2908 | | |
| PJ,PIT | 435 | 300 | .943 | .925 | 167 | .934 | .911 | 197 | .932 | .910 |
| PJ,JT,PIT | 420 | 244 | .954 | .936 | 144 | .943 | .921 | 163 | .944 | .923 |
| PJ,PJT,PIT | 360 | 139 | .974 | .958 | 92 | .964 | .941 | 104 | .964 | .943 |

As the $\hat{\Delta}$ and $\hat{\delta}$ values for the complete table are .954 and .939 for model {PJ,PIC}, .919 and .892 for {PJ,PTC} and .943 and .925 for {PJ,PIT}, we may conclude that the second- and higher-order interactions between J and the other variables are relatively unimportant. This conclusion is substantiated by the results of the subsamples. In each of the three analyses, the results of the two subsamples would lead to the same conclusions. This is an indication that the results are relatively stable. How (un)important the interactions with J are is difficult to say. The goodness of fit of the above model can certainly be improved by including parameters for interactions with J. However, even without such parameters the goodness of fit appears to be substantial. Therefore, the judgments seem sufficiently reliable.

## Discussion

In the preceding analyses we have refrained from the use of statistical tests because of two important reasons. Firstly, the use of the chi-square distribution for evaluating the fit of log-linear models assumes that the observations are sampled independently of each other. This assumption is violated in three ways: (a) the set of observations consists of subsets of acts that were generated by one and the same person, (b) the same set of acts was judged four times, i.e. by the four different judges, and (c) there are true and/or perceived sequential relationships among the acts (Bales, 1951). Secondly, the large number of acts and observations makes even small and uninteresting effects statistically significant. Therefore we prefer the use of descriptive indices like $\hat{\Delta}$ and $\hat{\delta}$ over the use of statistical tests. However, adequate tests for this kind of problem could be developed on the basis of Monte Carlo procedures.

In our case both the $\hat{\Delta}$ and $\hat{\delta}$ indices led to the same conclusion, i.e. that the qualitative judgments are sufficiently reliable. What is needed, however, is a set of guidelines that help to determine whether particular (high) values of $\hat{\Delta}$ and $\hat{\delta}$ denote 'sufficient', 'good', or 'perfect' reliability. Such guidelines might be developed through Monte Carlo simulations.

Finally, we would like to stress that even without the use of coefficients as $\hat{\Delta}$ and $\hat{\delta}$, log-linear analysis is a useful tool for studying inter-judge reliability because it shows the ways in which the judges' categorizations differ from each other.

References

Benzécri, J. P., and Collaborateurs. L'analyse des données: II L'analyse des correspondances. Paris: Dunod, 1973.

Bales, R.F. Interaction process analysis: A method for the study of small groups. Cambridge, Mass.: Addison -Wesley, 1950.

Bales, R. F. Some statistical problems in small group research. Journal of the American Statistical Association, 1951, 46, 311-322.

Bales, R. F. Personality and interpersonal behavior. New York: Holt, Rinehart, Winston, 1970.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. Discrete multivariate analysis. Cambridge, Mass.: MIT Press, 1975.

Bonett, D. G., and Bentler, P. M. Goodness-of-fit procedures for the evaluation of log-linear models. Psychological Bulletin, 1983, 93, 149-166.

Brown, M. B. Two-way and multiway frequency tables - Measures of association and the log-linear model (complete and incomplete tables). In Dixon, W. J. (Ed.), BMDP Statistical software. Berkeley: University of California Press, 1981.

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

Cohen, J. Weighted Kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin, 1968, 70, 213-220.

Cronbach, L. J., Gleser, G. C., and Rajaratnam, N. The dependability of Theory of generalizability for scores and profiles. New York: Wiley, 1972.

Everitt, B. S. The analysis of contingency tables. London: Chapman and Hall, 1977.

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

Gleser, G. C., Cronbach, L. J., and Rajaratnam, N. Generalizability scores influenced by multiple sources of variance. Psychometrika, 1965, 30, 395-418.

Guilford, J. P. Psychometric methods. New York: McGraw-Hill, 1954.

Landis, J. R., and Koch, G. G. A review of statistical methods in the analysis of data arising from observer reliability studies (Part I). Statistica Neerlandica, 1975, 29, 101-123.a

Landis, J. R., and Koch, G. G. A review of statistical methods in

the analysis of data arising from observer reliability studies (Part II). Statistica Neerlandica, 1975, 29, 151-161.b

Willemsen, T. M. Groups in discussion: A methodological and substantial study of interaction in small groups. Doctoral Dissertation, University of Leiden, The Netherlands, 1984.