Psychology Science, Volume 48, 2006 (1), p. 69-84

Exploring rater agreement: configurations of agreement and disagreement

ALEXANDER VON EYE[1], EUN-YOUNG MUN

Abstract

At the level of manifest categorical variables, a large number of coefficients and models for the examination of rater agreement has been proposed and used for descriptive and explanatory purposes. This article focuses on exploring rater agreement. Configural Frequency Analysis (CFA) is proposed as a method of exploration of cross-classifications of raters' judgements. CFA allows researchers to (1) examine individual cells and sets of cells in agreement tables; (2) examine cells that indicate disagreement; and (3) explore agreement and disagreement among three or more raters. Four CFA base models are discussed. The first is the model of rater agreement that is also used for Cohen's (1960) κ (kappa). This model proposes independence of raters' judgements. Deviations from this model suggest agreement or disagreement beyond chance. The second CFA model is based on a log-linear null model. This model is also used for Brennan and Prediger's (1981) κ_n. It proposes a uniform distribution of ratings. The third model is that of Tanner and Young (1985). This model proposes equal weights for agreement cases and independence otherwise. The fourth model is the quasi-independence model. This model allows one to blank out agreement cells and thus to focus solely on patterns of disagreement. Examples use data from applicant selection.

Key words: rater agreement; Configural Frequency Analysis (CFA); exploration; agreement types and antitypes; base models

[1] Please address correspondence concerning this article to Alexander von Eye, Michigan State University, Department of Psychology, 107 D Psychology Building, East Lansing, MI 48824-1116, USA

Exploring rater agreement: configurations of agreement and disagreement

Rater agreement is of great importance in many domains of research and life. Coders are expected to agree when coding video and audio tapes. The majority of voters agrees on which party will govern a country over the coming years. Olympic medals in gymnastics or boxing are awarded based on agreement among three or more judges, and so forth. For the analysis of rater agreement, many measures have been proposed, with Cohen's (1960) κ (kappa) and the coefficient of raw agreement being the most popular ones. In addition, an increasing number of models is being discussed that describe the variability in cross-tabulations of rater judgements, with log-linear models being prominent for manifest variable approaches (e.g., Tanner & Young, 1985; von Eye & Schuster, 2000) and latent class models being prominent for latent variables approaches (Schuster & Smith, 2002; Uebersax, 1993).

Both the coefficients and the models for rater agreement allow one to test hypotheses concerning a priori specified characteristics of rater agreement. For example, Cohen's κ is used to test the hypothesis that agreement is better than expected based on the chance model of rater independence. However, none of the coefficients or models proposed thus far embodies an exploratory component that would allow one to identify those rating categories in which two or more raters agree particularly strongly, those categories, in which they agree no better than chance, or those categories in which raters show more (or less) disagreement than expected based on chance. This article discusses Configural Frequency Analysis (CFA; Lienert & Krauth, 1975; von Eye, 2002; von Eye & Gutiérrez Peña, 2004) as a method of exploration of rater agreement.

This article consists of four parts. The first part introduces CFA as a method of exploration of rater agreement. The second part presents four models of CFA that are suitable for the exploration of rater agreement. The third part of this article presents data examples, and the article concludes with a discussion.

1. Exploring raters' judgements using Configural Frequency Analysis

In this article, we propose using Configural Frequency Analysis (CFA; Lienert & Krauth, 1975; von Eye, 2002) as a method for the exploration of cross-classifications of rater judgements. CFA is a method that allows one to examine individual cells of a cross-classification under the null hypothesis that $E[m_i] = E_i$, where $E[...]$ is the expectancy, m_i is the observed frequency in Cell i, E_i is the expected cell frequency, and i goes over all cells of the cross-classification (cf. duMouchel, 1999). CFA labels cells (also called *configurations*) as constituting *types* if $E[m_i] > E_i$. Configurations constitute *antitypes* if $E[m_i] < E_i$.

Three ways of determining estimated expected cell frequencies in CFA have been discussed (von Eye, 2002). The first involves estimating expected cell frequencies using a log-linear model. The second way involves using a priori probabilities, as for instance, in CFA of first differences between time-adjacent scores. The third way involves estimation based on distributional assumptions. A related approach, proposed by duMouchel (1999), uses empirical Bayesian methods of data exploration. DuMouchel's approach defines extreme cells only in terms of what would be a CFA type. Therefore, we propose using the methods of CFA which enable one to examine both cells with above expectancy and cells with below expectancy numbers of cases.

In the following sections, we focus on log-linear models for the estimation of expected cell frequencies, for two reasons. First, the majority of CFA models uses log-linear models to calculate expected cell frequencies. Second, Cohen's κ can be viewed as using a log-linear main effect base model, and Brennan and Prediger's κ_n can be viewed as using a log-linear no-effect base model, that is, a null model. In the context of CFA, the former model is termed a *First Order Base Model*, and the latter is termed a *Zero Order Base Model*.

CFA is typically performed in an exploratory context. In most applications, each cell in a cross-classification is examined with the goal of determining whether it constitutes a type or an antitype, or whether there is no significant discrepancy from expectation. A large number of tests has been proposed for this examination, including the z-test and the binomial test for multinomial sampling, and, when sampling is product-multinomial, Lehmacher's (1981) exact and approximative hypergeometric tests. Because the number of statistical decisions can be large, protection of the family-wise α is standard procedure in CFA applications (see Perli, Hommel, & Lehmacher, 1985).

When exploring the cross-classification of two raters, one has a number of options which cells to examine. The coefficient of raw agreement focuses on the cells in the main diagonal, also called the *agreement cells*. Cohen's (1960) κ and Brennan and Prediger's (1981) κ_n are *proportionate reduction in error* measures (see Fleiss, 1975). That is, these measures ask whether, overall, the number of observed instances of disagreement is below the expected number. This corresponds to asking whether the diagonal cells contain more cases of agreement than predicted from the base model. Thus, CFA of rater agreement can focus on the diagonal cells, looking, e.g., for agreement types. However, CFA can also be used to examine off-diagonal cells, looking, e.g., for disagreement antitypes. Alternatively, CFA can examine each cell in the table, looking for patterns of types and antitypes. In the following examples, we first opt for examining each cell, and focus on the specification of models on which the search for types and antitypes of agreement or disagreement can be based. Later, we examine selections of cells, for instance, the disagreement cells (cf. von Eye & von Eye, 2005).

Significant deviations from models that focus on all cells can suggest that the two raters

1) agree more often than expected, if $E[m_{ii}] > E_{ii}$ (agreement types);
2) agree less often than expected, if $E[m_{ii}] < E_{ii}$ (agreement antitypes);
3) disagree more often than expected, if $E[m_{ij}] > E_{ij}$ (for $i \neq j$) (disagreement types); or
4) disagree less often than expected, if $E[m_{ij}] < E_{ij}$ (for $i \neq j$) (disagreement antitypes).

This applies accordingly when agreement among more than two raters is explored. In the following section, we describe four CFA base models that we consider for the exploration of rater agreement.

2. CFA models for the exploration of rater agreement

The four CFA base models that are presented in this section, differ in the assumptions made concerning the agreement cells, that is, the cells in the diagonal of the cross-classification of two or more raters' judgements. The first two models, the ones for first order and for zero order CFA, do not make any particular assumption concerning the agreement cells, except that the frequency distribution in these cells follows the base models. For first order CFA, the base model is that of rater independence. For zero order CFA, the base model proposes that no effects exist. The third model considered here was proposed by Tanner and Young (1985) for explanatory analysis of agreement tables. This model proposes that raters place equal weights on the agreement cells. The fourth model is specific for the exploration of disagreement cells. This model blanks out the agreement cells and searches for types and antitypes in the disagreement cells. The following paragraphs describe these four models in more detail.

First or CFA of rater agreement. The base model of first order CFA was the original CFA model (Lienert, 1969). In the present context, it proposes independence among the *d* raters whose judgements are crossed. More specifically, the model proposes, for two raters,

$$\log m = \lambda 0 + \lambda_i^A + \lambda_j^B + e \, ,$$

where *m* is the vector of observed frequencies, λ_0 is the intercept, the λ_i^A are the parameters for the main effects of Rater A, the λ_i^B are the main effects for Rater B, and *e* is the vector of residuals. There are *J* main effect parameters ($J \geq 1$) for each rater, corresponding to the J + 1 rating categories used.

Instead of looking at the overall model fit, CFA examines individual cells and asks whether the above null hypothesis must be rejected. If a null hypothesis must be rejected, it suggests an agreement type or antitype, or a disagreement type or antitype. Each of these indicates a violation of the independence assumption and is interpreted individually. The typical result for agreement tables includes a number of agreement types and a number of disagreement antitypes. Data examples follow below.

Zero order CFA of rater agreement. Cohen's κ has been criticized for a number of reasons, two of which stand out and are discussed here. The first of the criticisms of Cohen's κ is known as marginal dependence. This characteristic indicates that if (1) the marginal probabilities are unequal and (2) at least one off-diagonal cell has a probability greater than zero, κ has an asymptotic maximum score of less than unity. As a result, a comparison of κ values across tables can be problematic. The second criticism, related to the first, is that κ can indicate low levels of agreement beyond chance although a vast proportion of judgements matches exactly. The reason for this characteristic is that large frequencies in diagonal cells can conform with expectation as specified in the main effect model, in particular if the marginals differ from each other.

To deal with these criticisms, Brennan and Prediger (1981) proposed using the uniform distribution model for a base model for κ instead of the main effect model of rater independence (cf. von Eye & Sörensen, 1991). The resulting measure of rater agreement, κ_n, does not suffer from these two criticized characteristics of Cohen's κ. For a comparative discussion of Cohen's (1960) κ and Brennan and Prediger's (1981) κ_n, see Hsu and Field (2003).

In the context of configural exploration of rater agreement, the same discussion can be carried. Types and antitypes from first order CFA share characteristics with κ. It can occur that the largest number of agreements does not stand out as a type, because this number conforms with the expectancy that is based on the model of variable independence. (This topic will be taken up again in the discussion section.)

Therefore, we suggest, in a fashion analogous to Brennan and Prediger's (1981) approach, also considering the null model for exploration of rater agreement. The model is log $m = \lambda$. Deviations from this model indicate that particular configurations of rating categories were observed more often (types) or less often (antitypes) than estimated by the null model. This implies that cells emerge as constituting types if they contain significantly more cases than the average cell, and cells emerge as constituting antitypes if they contain significantly fewer cases than the average cell.

Tanner and Young's (1985) equal weight agreement model as a CFA base model. Tanner and Young's (1985) equal weight agreement model (also called the *null-association agreement model*; see Schuster, 2002) assumes that the parameters for the interaction between Rater A and Rater B, λ_{ij}^{AB} , are all zero. In this respect, this model is identical to the base models for Cohen's κ and first order CFA. However, in addition, Tanner and Young's (1985) model, which is equivalent to Aickin's (1990) constant predictive probability model, posits an equal weight parameter for the diagonal cells, that is, the agreement cells. For two raters, the model can be formulated as the log-frequency model

$$\log m = \lambda_0 + \lambda_i^A + \lambda_J^B + \delta_{ij}\xi + e$$

where δ_{ij} is the vector that contains the weights and ξ is the parameter that is estimated for this vector, and A and B label the two raters. In Tanner and Young's model, $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$, else. For more than two raters, this model can be adapted in a straightforward way (von Eye & Mun, 2005). The expression exp(2ξ) is known to have a simple odds-ratio interpretation which reflects the degree of agreement. More specifically, the interpretation is

$$\exp(2\xi) = \frac{m_{ii}m_{jj}}{m_{ij}m_{ji}}$$

(Schuster, 2002). Thus, ξ can be compared to Cohen's κ.

In the present context, however, we are less interested in the overall degree of agreement. Instead, we ask whether, in particular cells, types and antitypes exist that contradict the equal weight agreement model. If such types and antitypes can be identified, they indicate local associations. Just as first order CFA of rater agreement, these associations suggest systematic patterns in the joint frequency distribution of the raters. In addition, however, they indicate that the hypothesis of equal weights in the agreement cells allows one to explain only part of the variation in the agreement table.

Quasi-independence model for the exploration of disagreement. In particular in contexts of rater training, it is important to know where raters disagree. Beyond chance disagreement may result in specific training or in re-specifications of agreement scales. A CFA base model that is suited to the exploration of disagreement cells is the log-linear quasi-independence model. For two raters, this model is

$$\log m = \lambda_0 + \lambda_i^A + \lambda_j^B + \sum_k \lambda_k + e$$

where the first three terms on the right hand side of the equation are the same as in the first order base model. The summation term describes the vectors needed to blank out the agreement cells. In the typical case, J such vectors are needed in a model with J being the number of rating categories, $i, j = 1, ..., J - 1$, and $k = 1, ..., J$. Types that result from this model indicate disagreement beyond chance, and antitypes indicate lack of disagreement that is beyond chance.

It is important to realize the difference between types and antitypes of disagreement that result from the model of quasi-independence and those that result from the first order CFA base model. Both models propose independence between raters. However, the first order CFA model estimates expected cell frequencies taking into account all cells, including the agreement cells. The quasi-independence base model estimates expected cell frequencies under exclusion of the agreement cells. Thus, types and antitypes of disagreement describe patterns of disagreement rather than judgements in general.

In the development of CFA as a statistical method, a model that is similar to the present quasi-independence model was discussed by Victor (1983) and Kieser and Victor (1999; cf. von Eye, 2002). Victor noted that routine CFA base models are specified to include all cells of a cross-classification. This includes those cells that end up being identified as types and antitypes. The existence of such types and antitypes can have two effects. First, cells that, otherwise, would be inconspicuous, can be turned into types or antitypes. Second, possibly existing types or antitypes can be obscured by other types or antitypes. In one word, the structure in a cross-classification can be misinterpreted because of the existence of the types and antitypes one is after. For an example of such a phenomenon see Kieser and Victor (1999).

The authors identify two major reasons for the obscuring of types and antitypes by other types and antitypes. The first reason is the well-known dependence of types and antitypes in a cross-classification. In extreme cases (typically in small tables), the existence of one type predetermines the existence of other types and antitypes (for a proof, see von Weber, Lautsch, & von Eye, 2003). The second reason lies in the routine CFA base model. This model assumes that types and antitypes do not exist. For these reasons, the authors proposed the above log-linear model of quasi-independence as a suitable CFA base model. In a first step, this model blanks out those cells that are expected to constitute types or antitypes. If this selection is valid, the remaining cells conform to the base model.

In the present context of exploration of disagreement cells, we propose a similar procedure. We propose blanking out the agreement cells, that is, the cells in the diagonal of an agreement table. The remaining cells, that is, the disagreement cells can then be analyzed under just any CFA base model, provided there are enough degrees of freedom left for the base model. A prime candidate for such a model is the original CFA base model of rater independence. However, other models are conceivable. For example, in training studies in which raters are trained to avoid scoring errors, one could test hypotheses concerning the location of errors over time.

3. Data examples

In the following paragraphs, we present data examples. We analyze data from a study on the agreement of raters on the qualification of job applicants in a large agency in the United States[2]. A sample of N = 465 interview protocols was examined by two evaluators. Each evaluator indicated on a four-point scale the degree to which an applicant was close to the profile specified in the advertisement of the position, with 1 indicating very good match and 4 indicating lack of match (see von Eye & Mun, 2005). Table 1 displays the observed frequency distribution of the cross-classification of the two evaluators.

Table 1 suggests that the two evaluators agree at a moderate level. Specifically, the Pearson $X^2 = 273.40$ indicates significant deviations from independence ($df = 9$; $p < 0.01$). Raw agreement is 54.41%, that is, the two evaluators agree in over 54% of their decisions exactly. For Cohen's κ, we calculate 0.364 ($se_\kappa = 0.032$; $p < 0.01$). This value suggests that the two evaluators agree to over 36% better than chance. For Brennan and Prediger's (1981) κ_n, we calculate 0.39. This value suggests that the two evaluators agree to almost 40% more often than expected based on the reference chance model that posits that both evaluators are independent and use the rating categories at equal rates. We now discuss the exploration of this cross-classification of raters' judgements.

First Order CFA. We now analyze the data in Table 1 under the four base models discussed in Section 2 of this article. We begin with First Order CFA. Sampling is considered multinomial. We select the z-test for the examination of the individual cells, with $\mathrm{z} = (\mathrm{m} - \mathrm{e})/\sqrt{\mathrm{e}}$ where m is the observed cell frequency and e is the estimated expected cell frequency. The significance level is set to $\alpha = 0.05$; after Bonferroni-adjustment, it is $\alpha^* = 0.05/16 = 0.003$. The chance model for First Order CFA is the log-linear main effect model $\log m = \lambda + \lambda_i^A + \lambda_j^B + e$, where the superscripts index the two raters. Table 2 displays the results of First Order CFA.

As was noted above, the overall goodness-of-fit Pearson $X^2 = 273.40$ ($df = 9$; $p < 0.01$) suggests significant deviations from the base model of rater independence. We thus can expect types and antitypes to emerge.

When testing for types and antitypes, the standard null hypothesis is $E[m_{ii}] = E_{ii}$. The standard alternative hypothesis is $E[m_{ii}] \neq E_{ii}$. Based on this specification of hypotheses, CFA tests are often considered two-sided, because the sign of the difference between the

Table 1:
Cross-classification of two evaluators' ratings of job interview protocols

		Evaluator B			
		1	2	3	4
Evaluator A	1	80	36	10	0
	2	30	67	41	2
	3	6	41	85	17
	4	0	4	25	21

[2] Thanks go to Neal Schmitt for making these data available.

Table 2:
First Order CFA of the rater agreement data in Table 1

Rater AB	*m*	*e*	*z*	*p*	
11	80	31.432	8.663	.000	Type
12	36	40.103	-.648	.518	
13	10	43.626	-5.091	.000	Antitype
14	0	10.839	-3.292	.002	Antitype
21	30	34.925	-.833	.404	
22	67	44.559	3.362	.000	Type
23	41	48.473	-1.073	.284	
24	2	12.043	-2.894	.004	
31	6	37.170	-5.113	.000	Antitype
32	41	47.424	-.933	.350	
33	85	51.589	4.652	.000	Type
34	17	12.817	1.168	.242	
41	0	12.473	-3.532	.000	Antitype
42	4	15.914	-2.987	.002	Antitype
43	25	17.312	1.848	.064	
44	21	4.301	8.052	.000	Type

observed and the estimated expected cell frequencies is not known to the testing procedures. The tail probability with which the protected α is compared is then $p = 2 \cdot \text{Prob}(Z > |z|)$, where Z is the test statistic estimated for a cell, and z is the critical value in the sampling distribution. However, many researchers who develop and use CFA, do take this sign into account. In a parallel fashion, the sign is taken into account in Bayesian data mining (see, e.g., DuMouchel, 1999). If the sign is taken into account, the tail probability is $p = \text{Prob}(Z > |z|)$ for types and $p = \text{Prob}(Z < |z|)$ for antitypes. While the former approach to significance testing in CFA is strict and accurate, the latter reflects (1) the data mining and exploratory characteristics of CFA application, (2) the common use of the method, and (3) the way CFA programs are written (see, e.g., the programs referred to in Krauth, 1993, or von Eye, 2001, 2002). In the remainder of this article, we apply, two-sided tests for the data examples.

Adopting this strategy, we find that first order CFA suggests the existence of four types and five antitypes. The types are constituted by Cells 1 1, 2 2, 3 3, and 4 4. These are the agreement cells, that is, the cells that contain the cases in which both raters showed perfect agreement. Each of these cells contains significantly more cases than could be expected based on chance.

The antitypes are constituted by Cells 1 3, 1 4, 3 1, 4 1, and 4 2. Each of these cells contains significantly fewer cases than could be expected based on chance. Interestingly, this pattern of antitypes suggests that instances in which these two raters disagree by two or more rating categories occur significantly less often than expected based on the independence model. A result of this kind cannot be found using such summary coefficients as Cohen's κ or Brennan and Prediger's κ_n. The same applies to the following observation.

None of the cells that indicate a difference of only one rating category between the two raters, constitutes a type or an antitype. These are Cells 1 2, 2 1, 2 3, 32, 3 4, and 4 3. Of these, only Cell 4 3 has a tendency to contain more cases than expected (p = 0.03). The Bonferroni procedure of protecting α prevents this cell from constituting a type.

Zero Order CFA. The second base model that we employ for the exploration of the data in Table 1 is Zero Order CFA. This analysis uses a chance model parallel to the one used for Brennan and Prediger's (1981) κ_n, that is, the log-linear null model log m = λ. Deviations from this model can be interpreted in a fashion analogous to deviations from the model for First Order CFA, the only difference being that whereas here, deviations indicate the existence of main effects, an interaction, or both, in First Order CFA, deviations only indicate the existence of an interaction, because main effects are taken into account. To create comparable results, Zero Order CFA was performed under the same specifications as First Order CFA for Table 2. Table 3 presents results.

The overall Pearson goodness-of-fit X^2 for the base model of Zero Order CFA of these data is 401.51 which indicates significant deviations from a uniform distribution (df = 15; p < 0.01). These deviations are realized in the form of a type-antitype pattern very similar to the one shown in Table 2. However, there are two differences of note. First, Cell 4 4 no longer constitutes a type. Thus, we cannot conclude that the two raters agree more often than expected from a no effect model in their ratings when they use Category 4. In fact, the observed frequency in Cell 4 4 is slightly below average. Second, most of the model-data discrepancies in Table 3 are larger than in Table 2. This difference reflects the differences between the base models: The main effect model used to estimate the expected cell frequencies in Table 2 takes more information into account than the no-effect model used for Table 3.

Table 3:
Zero Order CFA of the rater agreement data in Table 1

Configuration	*m*	*e*	*z*	*p*	
11	80	29.063	9.449	.000	Type
12	36	29.063	1.287	.198	
13	10	29.063	-3.536	.000	Antitype
14	0	29.063	-5.391	.000	Antitype
21	30	29.063	.174	.862	
22	67	29.063	7.037	.000	Type
23	41	29.063	2.214	.026	
24	2	29.063	-5.020	.000	Antitype
31	6	29.063	-4.278	.000	Antitype
32	41	29.063	2.214	.026	
33	85	29.063	10.376	.000	Type
34	17	29.063	-2.238	.026	
41	0	29.063	-5.391	.000	Antitype
42	4	29.063	-4.649	.000	Antitype
43	25	29.063	-.754	.452	
44	21	29.063	-1.496	.134	

The equal weight agreement base model. We now employ Tanner and Young's (1985) equal weight agreement model for the exploration of the agreement data in Table 1. As was indicated above, this model goes beyond the first order CFA model by including a parameter for the agreement cells. This parameter can be interpreted as an indicator of strength of agreement. Performing a CFA when this parameter is included, implies asking whether types and antitypes of agreement or disagreement exist once we take into account the strength of agreement between the two raters. Table 4 displays the results from this analysis, for which we also used the *z*-test, and the Bonferroni procedure to protect α.

Table 4:
CFA of the agreement data in Table 1 under the equal weight agreement base model

Configuration	*m*	*e*	*z*	*p*	
11	80	64.389	1.945	.052	
12	36	25.559	2.065	.039	
13	10	27.454	-3.331	.000	Antitype
14	0	8.598	-2.932	.003	Antitype
21	30	20.461	2.109	.035	
22	67	82.931	-1.749	.080	
23	41	27.877	2.485	.013	
24	2	8.731	-2.278	.023	
31	6	21.139	-3.293	.000	Antitype
32	41	26.812	2.740	.006	
33	85	92.029	-.733	.463	
34	17	9.020	2.657	.008	
41	0	10.011	-3.164	.002	Antitype
42	4	12.698	-2.441	.015	
43	25	13.640	3.076	.002	Type
44	21	13.651	1.989	.047	

The equal weight agreement base model for the data in Tables 1 and 4 suggests significant data-model discrepancies (X^2 = 101.95; *df* = 8; $p < 0.01$). This model is significantly better than the main effect model without the equal weight agreement parameter that was used for the analyses in Table 2 (ΔX^2 = 171.44; Δdf = 1; $p < 0.01$). Still, the model itself must be rejected. Therefore, we cannot interpret the equal weight agreement parameter (ξ = 1.16; se = 0.09), and we expect types and antitypes to emerge. Table 4 shows that 1 type and 4 antitypes emerge when strength of agreement is taken into account.

The pattern of types and antitypes differs greatly from the patterns in Tables 2 and 3. None of the agreement cells constitutes a type anymore. Taking into account strength of agreement thus allows one to explain that part of the variability in the agreement table that is due to exactly matching judgements. However, there are disagreement antitypes, and, for the first time in the present analyses, disagreement types. The antitypes overlap with the antitypes identified by First Order CFA. They indicate that disagreement by more than one scale point is less likely than expected based on the assumption of rater independence, both with and without taking into account strength of agreement. The sole type is constituted by Cell 4 3. It suggests that disagreement by just one scale point is more likely than expected when strength of agreement is taken into account. Cells 1 2, 2 1, 2 3, 3 2, and 3 4 support this interpretation, although none of these indicates beyond chance agreement that is significant after the Bonferroni adjustment. Probability poolers as discussed by Darlington and Hayes (1999) and von Eye (2002) can be used to test the hypothesis that disagreement by just one scale point is more likely than expected by the equal weight agreement base model.

Log-linear models of quasi-independence for the analysis of disagreement. We now turn to the exploration of patterns of disagreement (cf. von Eye & von Eye, 2005). To explore disagreement, we employ the log-linear model of quasi-independence and blank out the agreement cells in the main diagonal of the agreement table. The CFA results under this base model appear in Table 5. To make the results comparable, we again used the *z*-test and the Bonferroni procedure.

The overall Pearson goodness-of-fit for the base model in Table 5 shows significant data-model discrepancies ($X^2 = 77.73$; $df = 5$; $p < 0.01$). We thus conclude that the disagreement cells do not follow a distribution that is conform with the hypothesis of rater independence. CFA identifies two types, constituted by Cells 3 4, and 4 3. These types suggest that disagreement by one scale point is more likely than expected under a model of independence that focuses exclusively on the disagreement cells. Using probability poolers such as Stouffer's *Z*, one can test the hypotheses whether including Cells 1 2, 2 3, and 3 2 in a group that also contains the type-constituting cells 3 4, and 4 3 (as well as Cell 2 1) supports the state-

Table 5:
CFA of the data in Table 1 under the log-linear quasi-independence model, blanking out the agreement cells

Configuration	*m*	*e*	*z*	*p*	
11	80	80.000	.000	-	
12	36	23.571	2.560	.010	
13	10	18.847	-2.038	.041	
14	0	3.582	-1.893	.059	
21	30	18.049	2.813	.005	
22	67	67.000	.000	-	
23	41	46.174	-.761	.446	
24	2	8.776	-2.287	.022	
31	6	13.659	-2.072	.038	
32	41	43.700	-.408	.683	
33	85	85.000	.000	-	
34	17	6.641	4.019	.000	Type
41	0	4.292	-2.072	.038	
42	4	13.730	-2.626	.008	
43	25	10.979	4.232	.000	Type
44	21	21.000	.000	-	

ment that, overall, disagreement by one scale point is more likely than expected. Stouffer's Z is estimated by

$$Z = \sum_{i=1}^{t} Z_i / \sqrt{t} \ ,$$

where the z_i are the z scores that correspond to the probabilities of the configurations that are being pooled, and t is the number of configurations being pooled. For the present data, we calculate $Z = 12.45/\sqrt{6} = 5.085$ and $p < 0.01$. We thus consider this statement supported.

4. Summary and discussion

This article proposed using Configural Frequency Analysis for the exploration of agreement tables. The exploration of such tables is of interest whenever researchers (a) need to know more about patterns of agreement and disagreement than can be provided by such coefficients as Cohen's κ or raw agreement, and (b) hypotheses specific enough to formulate models do not exist. For example, when raters or coders are trained, particular tendencies to disagree can be detected and training can focus on avoiding such patterns. These tendencies may be unknown before analysis. Also, trends shown by raters can be detected using the configural approach.

4.1 Comparing the models proposed for exploration of rater agreement

This article proposes four log-linear models as base models for the configural exploration of rater agreement and disagreement. These models are sensitive to different data characteristics and thus can suggest different appraisals of agreement and disagreement. We now discuss differences among these four models, and provide suggestions as to when each of the four models is most suitable.

The first base model discussed here is that of rater independence. This model takes the rater main effects into account. In different words, this base model takes into account the possibly differing rates with which raters use rating categories. As a consequence, types and antitypes of agreement and disagreement do not reflect the number of occurrences of a judgement pattern. Instead, they reflect the number of occurrences beyond (or below) what could be expected considering the rates with which raters use the rating categories. Types and antitypes from this base model therefore reflect local interactions that result in deviations from the probability pattern that is conform with the main effects model.

One implication of these characteristics of the main effect model is that large cell frequencies can be conform with the main effect model and may thus not result in types. Consider the artificial data in the 2 x 2 cross-classification in Table 6. In this table, the cell with the largest frequency fails to constitute an agreement type because the deviation from the main effect model is non-significant. Thus, although agreement is very high (raw agreement is 90.18%), the main effect base model of first order CFA does not label Cell 1 1 as out of

Table 6:
Sample table in which the cell with the largest frequency does not constitute a type when analyzed with first order CFA

Cell index	Cell frequencies		Statistical tests	
	observed	expected	z	$p(z)$
11	94	88.39	.59	.56
12	5	10.61	-1.72	.08
21	6	11.61	-1.65	.10
22	7	1.39	4.75	< .01

the expected, although it carries the vast majority (93.07%) of the agreement cases. We analyze the data in Table 6 using the z-test and the Bonferroni procedure. The adjusted significance threshold is $\alpha^* = 0.0125$.

The results in Table 6 show that both raters made extensive use of the first rating category. In 94 out of 112 cases, both raters used this category. In 7 additional cases, they both used Category 2. However, Cell 1 1 does not constitute an agreement type. Instead, Cell 2 2 constitutes an agreement type although it shows no more than 7 cases. Thus, this example illustrates that agreement and disagreement types and antitypes reflect interactions rather than the mere magnitude of a cell probability. Cohen's κ for the example in Table 6 is 0.505.

If researchers wish that agreement and disagreement types and antitypes also reflect the magnitude of a cell probability, the null model of no effects may be the appropriate base model. A re-analysis of the data in Table 6, presented in Table 7, illustrates the differential characteristics of the First Order and the Zero Order CFA base models.

Zero Order CFA identifies Cell 1 1 as constituting an agreement type, Cells 1 2 and 2 1 as constituting disagreement antitypes, and Cell 2 2 as constituting an agreement antitype. In different words, for each cell in this table, Zero Order CFA suggests a different type/antitype decision than First Order CFA. The probability of Cell 1 1 is significantly above average, and the probabilities of the other three cells are significantly below average. Both the main effects and the interaction in this table are responsible for this result. As was indicated above, the choice of Brennan and Prediger's (1981) coefficient κ_n over Cohen's κ (1960) can use the same arguments as the choice of the Zero Order CFA base model over the base model of First order CFA.

Table 7:
Zero Order CFA of the data in Table 6

Cell index	Cell frequencies		Statistical tests	
	observed	expected	z	$p(z)$
11	94	28	12.47	< 0.01
12	5	28	-4.35	< 0.01
21	6	28	-4.16	< 0.01
22	7	28	-3.97	< 0.01

The third base model that is proposed here is the equal weight agreement model of Tanner and Young (1985). This model involves a parameter that can be interpreted as a measure of strength of agreement. Types and antitypes of agreement and disagreement therefore reflect local interactions that go beyond what can be explained based on the rater main effects and knowledge that describes the strength of agreement.

When comparing this base model with the fourth of the base models proposed here, it is of importance to note that the Tanner and Young model is based on all cells in a table, and that statements resulting from this analysis concern both agreement and disagreement patterns. The fourth model proposed here focuses on disagreement, at the exclusion of the agreement cells. Therefore, the expected probabilities reflect only the structure of disagreement. The possible existence of types and antitypes of agreement can no longer obscure possible types and antitypes of disagreement. Information concerning agreement is not taken into account. The interpretation of types and antitypes of disagreement rests on the following assumption. If the covariation of two or more raters is carried by their agreement, the distribution in the disagreement cells should be random. Therefore, neither disagreement types nor antitypes should surface. If, however, such types and antitypes surface, systematic patterns of disagreement exist that may be worth further consideration, for example in training programs.

4.2 Alternative base models

Other base models for the exploration of rater agreement and disagreement can be considered. For example, one can devise a model that takes into account the equal weight agreement parameter, but not the main effects. The resulting types and antitypes would then reflect only deviations from strength of agreement as viewed from a null model. Here again, both main effects and local interactions can be responsible for emerging types and antitypes. However, before propagating this model, the interpretation of the equal weight agreement parameter under the null model needs to be specified in more detail.

Another model that may be of interest is the Zero Order CFA base model, applied to the agreement table when the agreement cells are blanked out. This model allows one to examine the disagreement cells with the average disagreement probability in mind. The difference between this model and the fourth model proposed here can be described using the same arguments as the description of the difference between the first two models proposed here.

The models proposed here can be extended in a natural way. First, more than two raters can be analyzed simultaneously. Second, classification and grouping variables can be taken into account. One can ask, for example, whether agreement patterns are the same in samples of female and male raters. Two-sample CFA allows one to answer this question. Third, covariates can be taken into account. One can ask whether types and antitypes of agreement and disagreement still exist when knowledge about other variables is considered. Finally, models that take into account scale characteristics, e.g., Goodman's linear-by-linear association model, can be placed in the context of exploration of agreement tables.

From a computational perspective, it can be noted that all the models discussed in this article can be estimated using general purpose statistical packages such as SAS, SYSTAT, or SPSS, or more specialized software such as Lem. Some of the CFA software (von Eye, 2001) is also capable of estimating these models.

In sum, this article proposes tools for the exploratory analysis of agreement tables. Coefficients of agreement provide information about various aspects of strength of agreement. The methods proposed here allow one to identify those patterns of agreement and disagreement that deviate in particular from assumptions that are specified in the form of base models. These assumptions concern either the entire table or a selection of cells.

References

1. Aickin, M. (1990). Maximum likelihood estimation of agreement in the constant predictive probability model and its relation to Cohen's kappa. Biometrics, 46, 293 - 302.
2. Brennan, R.L., & Prediger, D.J. (1981). Coefficient kappa: some uses, misuses, and alternatives. Educational and Psychological Measurement, 41, 687 - 699.
3. Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20, 37 - 46.
4. Darlington, R.B., & Hayes, A.F. (2000). Combining independent p values: Extensions of the Stouffer and binomial methods. Psychological Methods, 5, 496 - 515.
5. DuMouchel, W. (1999). Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system. The American Statistician, 53, 177 - 190.
6. Fleiss, J.L. (1975). Measuring agreement between two judges in the presence or absence of a trait. Biometrics, 31, 651 - 659.
7. Hsu, L.M., & Field, R. (2003). Interrater agreement measures: Comments on Kappa$_n$, Cohen's Kappa, Scott's π, and Aickin's α. Understanding Statistics, 2, 205 - 219.
8. Kieser, M., & Victor, N. (1999). Configural Frequency Analysis (CFA) revisited - A new look at an old approach. Biometrical Journal, 41, 967 - 983.
9. Krauth, J. (1993). Einführung in die Konfigurationsfrequenzanalyse (KFA). Weinheim: Beltz, Psychologie Verlags Union.
10. Lehmacher, W. (1981). A more powerful simultaneous test procedure in Configural Frequency Analysis. Biometrical Journal, 23, 429 - 436.
11. Lienert, G.A. (1969). Die Konfigurationsfrequenzanalyse als Klassifikationsmittel in der klinischen Psychologie, In: M. Irle (Hrsg.), Bericht über den 26. Kongreß der Deutschen Gesellschaft für Psychologie 1968 in Tübingen (pp. 244 - 255). Göttingen: Hogrefe.
12. Lienert, G.A., & Krauth, J. (1975). Configural frequency analysis as a statistical tool for defining types. Educational and Psychological Measurement, 35, 231 - 238.
13. Perli, H.-G., Hommel, G., & Lehmacher, W. (1985). Sequentially rejective test procedures for detecting outlying cells in one- and two-sample multinomial experiments. Biometrical Journal, 27, 885 - 893.
14. Schuster, C. (2002). A mixture model approach to indexing rater agreement. British Journal of Mathematical and Statistical Psychology, 55, 289 - 303.
15. Schuster, C., & Smith, D.A. (2002). Indexing systematic rater agreement with a latent class model. Psychological Methods, 7, 384 - 395.
16. Tanner, M. A., & Young, M.A. (1985). Modeling agreement among raters. Journal of the American Statistical Association, 80, 175 - 180.
17. Uebersax, J.S. (1993). Statistical modeling of expert ratings on medical treatment appropriateness. Journal of the American Statistical Association, 88, 421 - 427.
18. Victor, N. (1989). An alternative approach to configural frequency analysis. Methodika, 3, 61-73.

19. von Eye, A. (2001). Configural Frequency Analysis - A program for 32 bit Windows operating systems. Manual for program Version 2000. Methods of Psychological Research - Online, 6, 140.
20. von Eye, A. (2002). Configural Frequency Analysis - Methods, Models, and Applications. Mahwah, NJ: Lawrence Erlbaum.
21. von Eye, A., & Gutiérrez Peña, E. (2004). Configural Frequency Analysis - the search for extreme cells. Journal of Applied Statistics, 31, 981 - 997.
22. von Eye, A., & Mun, E.Y. (2005). Analyzing rater agreement - manifest variable models. Mahwah, NJ: Lawrence Erlbaum.
23. von Eye, A., & Brandtstädter, J. (1988). Application of prediction analysis to cross-classifications of ordinal data. Biometrical Journal, 30, 651-655.
24. von Eye, A., & Schuster, C. (2000). Log-linear models for rater agreement. Multiciência, 4, 38 - 56.
25. von Eye, A., & Sörensen, S. (1991). Models of chance when measuring interrater agreement with kappa. Biometrical Journal, 33, 781-787.
26. von Eye, A., & von Eye, M. (2005). Can One Use Cohen's Kappa to Examine Disagreement? Methodology, 1, 129-142.
27. von Weber, S., Lautsch, E., & von Eye, A. (2003). At the limits of Configural Frequency Analysis. Psychology Science. 45, 339-345.