Significance Tests for the Measure of Raw Agreement

von Eye, Alexander; Mair, Patrick; Schauerhuber, Michael

Published: 01/01/2006

Document Version
Publisher's PDF, also known as Version of record

Significance Tests for the Measure of Raw Agreement

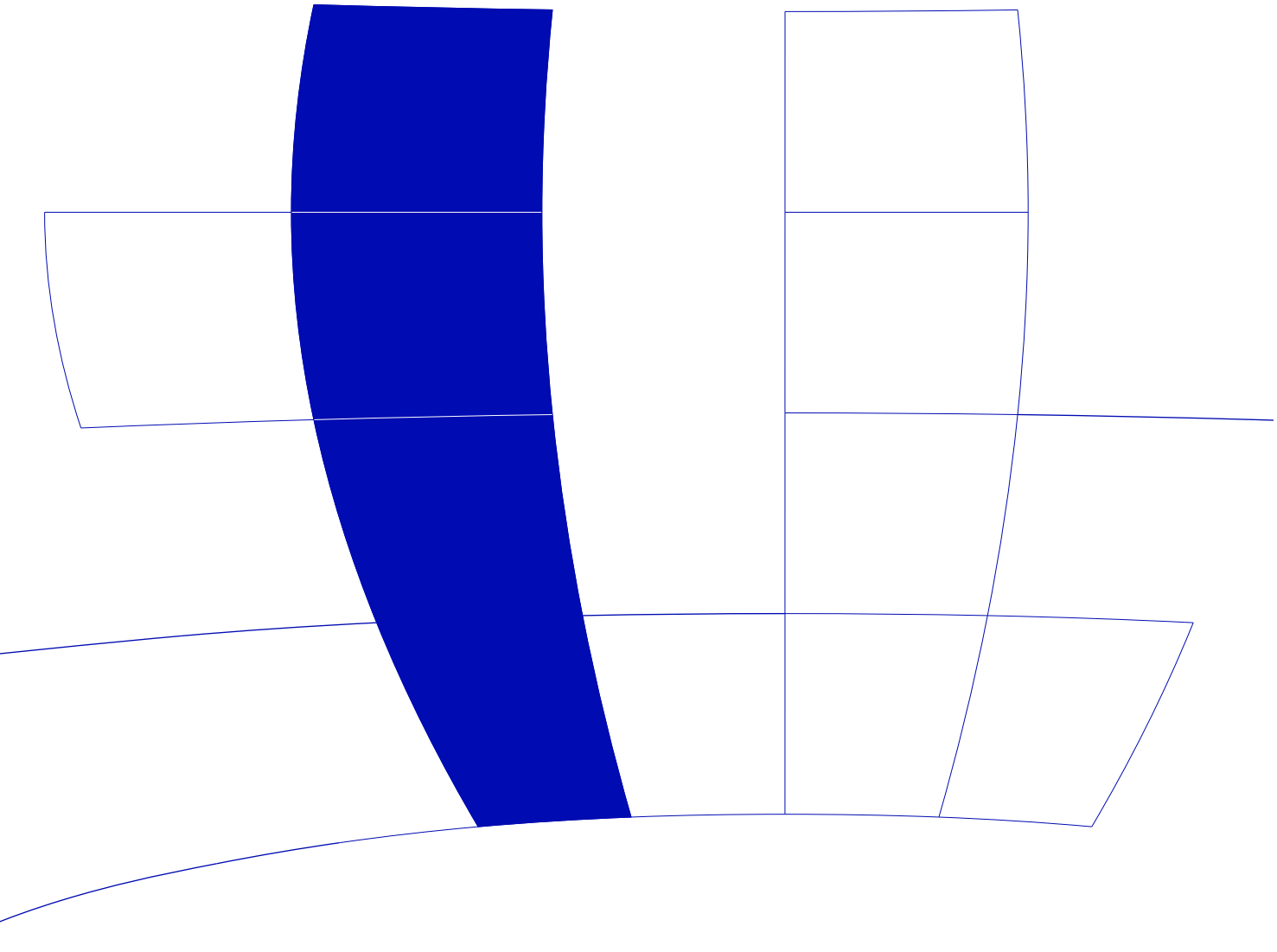
Alexander von Eye, Patrick Mair, Michael Schauerhuber

Department of Statistics and Mathematics
Wirtschaftsuniversität Wien

Research Report Series

Report 42
November 2006

http://statmath.wu-wien.ac.at/



Significance Tests for the Measure of Raw Agreement

Alexander von Eye, Patrick Mair, Michael Schauerhuber

November 2006

Abstract

Significance tests for the measure of raw agreement are proposed. First, it is shown that the measure of raw agreement can be expressed as a proportionate reduction-in-error measure, sharing this characteristic with Cohen's κ and Brennan and Prediger's κ_n. Second, it is shown that the coefficient of raw agreement is linearly related to Brennan and Prediger's κ_n. Therefore, using the same base model for the estimation of expected cell frequencies as Brennan and Prediger's κ_n, one can devise significance tests for the measure of raw agreement. Two tests are proposed. The first uses Stouffer's Z, a probability pooler. The second test is the binomial test. A data example analyzes the agreement between two psychiatrists' diagnoses. The covariance structure of the agreement cells in a rater by rater table is described. Simulation studies show the performance and power functions of the test statistics.

1 Introduction

Three approaches to assessing agreement among raters have been proposed [1]. The first approach is the most popular one. It involves calculating simple coefficients of rater agreement such as Cohen's κ (kappa) [2], Brennan and Prediger's κ_n [3], or the measure of raw agreement. κ is an index of agreement beyond chance. κ_n is an index of agreement beyond what is predicted by a null model for the agreement table. The measure of raw agreement indicates the number of instances in which raters agree exactly, relative to the total number of judgements. The second approach involves testing manifest variable models of rater agreement. These models represent hypotheses concerning the structure of agreement tables, that is, cross-classifications of raters' judgements. Examples of such models include Tanner and Young's equal weight agreement model [4] which proposes that the instances of agreement carry equal weight across all rating categories. This model includes a parameter that can be interpreted as an index of strength of agreement. The third approach involves specifying latent variable models. Examples of such models include Uebersax and Grove's [5] and Schuster's [6] latent variable and mixture models.
The models as well as the coefficients of rater agreement come with significance tests which allow one to test null hypotheses that, for example, agreement is no better than chance. There is only one exception, the coefficient of raw agreement. This coefficient is typically reported as a descriptive measure only. Null hypotheses are not specified and significance tests are not reported. In this article, a null hypothesis for the coefficient of raw agreement is proposed, two significance tests are described, and their characteristics are examined.

2 The Coefficient of Raw Agreement

The coefficient of raw agreement, ra, is a measure of the proportion of instances of agreement. Consider an I x I cross-classification of two raters' judgements. Each rater uses an I-category scale to evaluate N objects or events. The coefficient ra is defined as the proportion of exact agreements. To specify ra, we first determine the sum $\theta_1 = \sum_{i=1}^{I} p_{ii}$, that is, the probability of all

instances in which the two raters agree exactly, where p_{ii} is the probability of cases in agreement cell ii. θ_1 is estimated by $\hat{\theta}_1 = \frac{1}{N}\sum_i n_{ii}$, where n_{ii} is the observed frequency in cell ii. We define the coefficient of raw agreement to be $ra = \theta_1$. The estimate of the coefficient of raw agreement, $\hat{ra}$, indicates the proportion of the total number of judgements that are made in full agreement. It is $\hat{ra} = \hat{\theta}_1$. The coefficient $\hat{ra}$ is typically presented as a proportion or a percentage of agreement cases. Generalizations to more than two raters are straightforward.

Data example: The following data example re-analyzes data presented by von Eye and Mun [1], and von Eye and Schuster [7]. The data describe results from a study on the reliability of psychiatric diagnoses. Two psychiatrists re-evaluated the files of $N = 129$ patients. The patients had at intake been diagnosed as clinically depressed. The psychiatrists evaluated the severity of the patients' depression using the rating categories 1 = not depressed, 2 = mildly depressed, and 3 = clinically depressed. Table 2 presents the cross-classification of the two psychiatrists' diagnoses. In the following analysis, we ask what the proportion of diagnoses is that match exactly. In addition, we calculate Cohen's κ to also provide information of agreement beyond chance.

		Psychiatrist 2: Severity of Depression			
		1	2	3	Row Sums
	1	11	2	19	32
		2.98	*3.22*	*25.80*	
Psychiatrist 1:	2	1	3	3	7
Severity of Depression		*0.65*	*0.71*	*5.64*	
	3	0	8	82	90
		8.37	*9.07*	*72.56*	
Column Sums		12	13	104	N=129

Table 1: Two psychiatrists' depression diagnoses (estimated expected cell frequencies in *italics*)

The proportion of raw agreement for the two psychiatrists' diagnoses is $\hat{ra} = 96/129 = 0.74$. In words, we find that the two psychiatrists agree in almost three quarters of their diagnoses exactly. Table 2 also displays the expected cell frequencies for Cohen's κ. (The formula for κ follows in section 3, below.) We calculate $\hat{\kappa} = 0.38$ and $\hat{se}_\kappa = 0.079$ ($z = 4.747$; $p < 0.01$). These values indicate that the proportion of raw agreement of 0.74 is better by 38% than what one would expect based on chance alone. This difference to chance agreement is significant.
We now ask whether significance tests for ra can be formulated. To specify a significance test for the coefficient of raw agreement, we need to answer the question as to what to test against. To answer this question, we proceed in two steps. First, we express three measures of rater agreement in terms of measures of proportionate reduction in error (Fleiss, [8]), Cohen's κ [2], Brennan and Prediger's κ_n [3], and the measure of raw agreement. Second, we show that the measure of raw agreement is a linear transformation of Brennan and Prediger's κ_n, but not Cohen's κ. Therefore, we use the same base model as for Brennan and Prediger's measure to estimate expected cell frequencies. Using these frequencies, a null hypothesis for the measure of raw agreement can be specified, and significance tests can be proposed.

3 Three Coefficients of Rater Agreement

The best known measure of rater agreement is Cohen's κ. This coefficient uses a base model that takes into account the rates with which raters use the rating categories. For two raters, the coefficient is

$$\kappa = \frac{\sum_i p_{ii} - \sum_i p_{i.}\ p_{.i}}{1 - \sum_i p_{i.}\ p_{.i}}$$

and it is estimated by

$$\hat{\kappa} \;=\; \frac{N \sum_i m_{ii} \;-\; \sum_i m_{i.}\; m_{.i}}{N^2 \;-\; \sum_i m_{i.}\; m_{.i}}$$

where $p_{i.}$ and $p_{.i}$ are the row and the column probabilities, and $m_{i.}$ and $m_{.i}$ are the row and the column marginal frequencies, respectively. Obviously, the first term in the numerator of κ contains the quantity θ_1 which is also used for the coefficient of raw agreement. The second term, $\theta_2 = \Sigma_{\mathrm{i}}\; p_{i.} p_{.i}$, indicates the probability of cases in Cell ii that is expected based on the assumption of independence of the two raters. It takes the rates into account with which the raters use the rating categories. These rates are $p_{i.}$ and $p_{.i}$. The first term in the denominator contains the maximum probability of agreement cases and the second term in the denominator is, again, θ_2. Using the quantities θ_1 and θ_2, we can express κ as

$$\kappa \;=\; \frac{\theta_1 \;-\; \theta_2}{1 \;-\; \theta_2}$$

The third coefficient of rater agreement that we use in the present article is Brennan and Prediger's (1981) κ_n [3]. This measure is defined as

$$\kappa_n \;=\; \frac{\theta_1 \;-\; \frac{1}{I}}{1 \;-\; \frac{1}{I}}$$

where I is the number of rows and columns of the agreement table. The coefficient κ_n can be derived from the hypothesis that $p_{\mathrm{i}.} = p_{.\mathrm{i}} = 1/I$. This topic will be taken up again below. κ_n differs from κ in that it does not take into account the rates with which the raters use the I rating categories. Discussions of κ and κ_n can be found in Hsu and Field [9], von Eye and Mun [1], and von Eye and Sörensen [10].

To compare ra with κ and κ_{n}, we specify a general proportionate reduction-in-error measure of rater agreement,

$$\kappa_g \;=\; \frac{\theta_1 \;-\; c}{1 \;-\; c}$$

and we find that, for Cohen's κ, $c = \theta_2$, for Brennan and Prediger's κ_n, $c = 1/I$, and for ra, $c = 0$. We thus conclude that the difference between these three measures lies in the selection of a reference for θ_1, the proportion of incidences of exact agreement. For Cohen's κ, this reference is θ_2, the proportion of exact matches that is estimated under a log-linear main effect model, that is, the model of rater independence. This model takes the rates into account with which the raters use each of the I rating categories. For Brennan and Prediger's κ_n, the reference is the proportion of cells in which exact matches can be found. This proportion depends on the size of the agreement table, but not on the marginal distributions of ratings. The choice of Brennan and Prediger's reference corresponds to estimating the proportion of expected exact matches based on a log-linear null model. For the coefficient of raw agreement, the reference is zero. That is, the coefficient of raw agreement provides information about how large the proportion of exact matches is compared to no exact matches.

Looking at the three measures, κ, κ_n, and ra from this perspective, allows one to notice that, for a given table size and θ_1, Cohen's κ can vary widely, depending on θ_2 which typically is estimated from the data. In contrast, for both κ_n and ra, c does not depend on the data. Indeed, a closer comparison of κ_n and ra shows that the two measures are linear transformations of each other. We find that

$$\kappa_n \;=\; -\; \frac{1}{I \;-\; 1} \;+\; ra\; (\frac{I}{I \;-\; 1})$$

where I is, as before, the number of rating categories.

To illustrate, consider Figure 1. This figure displays the relationship between the general measure of agreement defined above and the quantities θ_1 and θ_2 for $0 \leq \theta_1, \theta_2 \leq 1$. Figure 1 shows clearly that κ varies depending on θ_2. Setting θ_2 constant, as is done for κ_n and ra, yields a straight-line relationship between κ_g and θ_1. Figure 2 shows this for $c = 0.3$. Figure 3 shows this for $c = 0$.

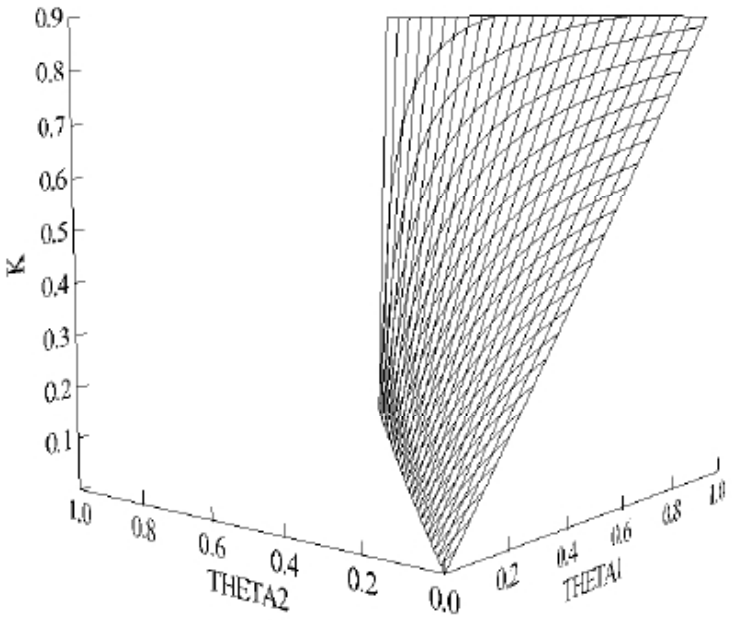


Figure 1: Relationship between the general measure of agreement, κ_g, and the quantities θ_1 and θ_2

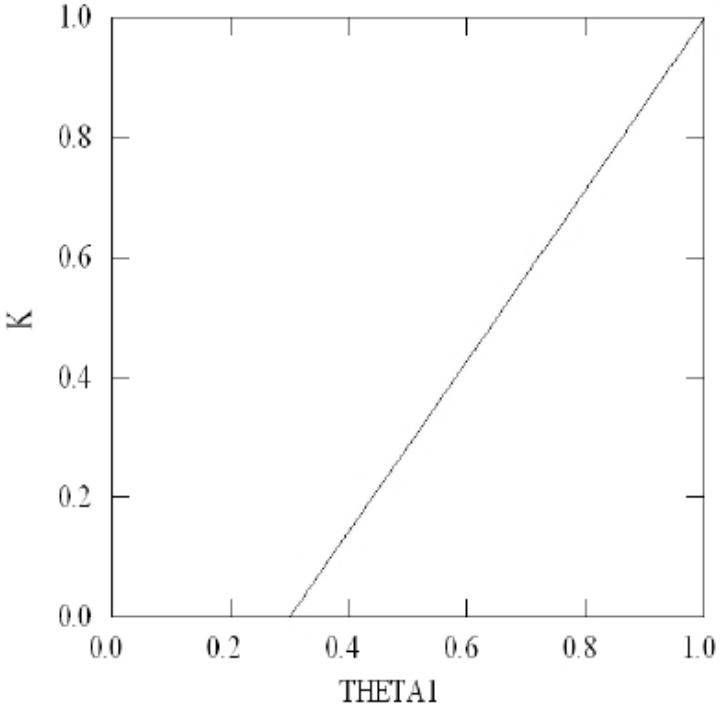


Figure 2: κ_g for $c = 0.3$

The relationship between κ_n and ra is displayed in Figure 4. Figure 4 shows, that, for any given I, the relationship between ra and κ_n is linear. Therefore, the magnitude of the constant

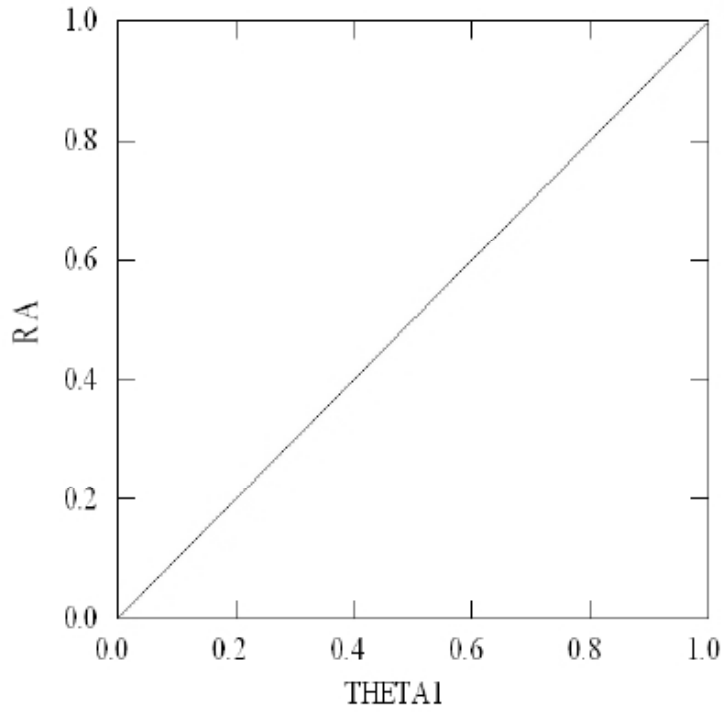


Figure 3: κ_g for $c = 0$

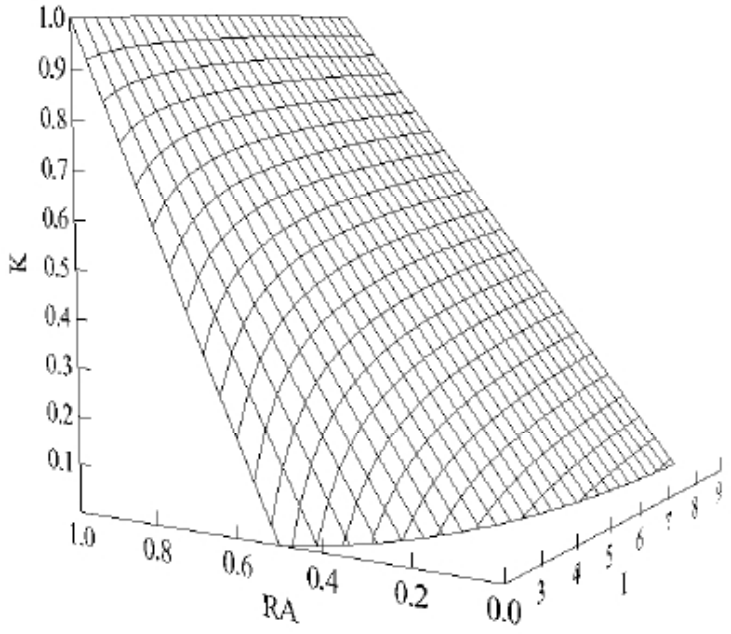


Figure 4: Relationship between κ_n, ra and I

c is of no importance beyond scaling, as long as $0 \leq c < 1$. In addition, because of this linear

relationship, any significance test that is valid for κ_n is also valid for ra and vice versa. This does not apply to significance tests for Cohen's κ because the relation between κ, κ_n, and ra is nonlinear.

4 Two Significance Tests for the Coefficient of Raw Agreement

In this section, we propose two significance tests for the coefficient of raw agreement. The first test uses Stouffer's Z [11], a well known probability pooler. The second test is the exact binomial test.

4.1 Stouffer's Z as a Significance Test for the Coefficient of Raw Agreement

For the following considerations, let m_{ii} be the expected frequency in Cell ii, that is, in an agreement cell, and $\hat{m}_{ii}$ the estimated expected frequency for Cell ii. Furthermore, let $\hat{m}_{ii}$ be estimated under the log-frequency model $\log m = \lambda_0 + e$, where λ_0 is the constant and e is the residual vector. This model is used as the base model for Brennan and Prediger's coefficient, κ_n. It implies a uniform frequency distribution. The residual in Cell ii is estimated as

$$z_{ii} = \frac{m_{ii} - N/I^2}{\sqrt{N/I^2}}$$

The sum of these z_{ii}-scores can be used as the test statistic for raw agreement. The test statistic

$$Z = \frac{1}{\sqrt{I}} \sum_i z_{ii}$$

with $i = 1, ..., I$, is known as Stouffer's Z. This statistic is approximately $N(0; 1)$ distributed (see also Darlington & Hayes [12] and Strube [13]). Z can be used to test the hypothesis that the portion of exact agreement is greater than $1/I$, that is, H_1: $\theta_1 > 1/I$. The term 1/I results from the assumption that we have a uniformly distributed probability distribution. Thus, under this model the probability for a cell count is $1/\text{I}^2$. Since the raw agreement is the sum of the probabilities of the main diagonal $\hat{\theta}_1 = I\ (1\ /\ I^2) = 1\ /\ I$, a feasible formulation of the null hypothesis ("agreement no better than chance") is H_0: $\theta_1 = 1/I$.

4.2 The Binomial Test for the Coefficient of Raw Agreement

The binomial test described in this section is of use when an exact test is desired. This occurs when the number of observations is small and the asymptotic result of a normally distributed Z is not trusted. The test is based on just one parameter, the probability of the event under study, p. In the present context, this is the overall probability of perfect matches in raters' judgements. Here, in the discussion of significance tests for the coefficient of raw agreement, we estimate p using the same log-frequency model as for Stouffer's Z [11] or Brennan and Prediger's κ_n [3], that is, $\log m = \lambda_0 + e$. We obtain the estimator $\hat{p} = 1/I$. This results again from the assumption that we have a uniform probability distribution under H_0 (cf. $\hat{\theta}_1$ in the previous section). The one-sided tail probability for the observed total number of perfect matches is

$$P = \sum_{j=\sum m_{ii}}^{N} \binom{N}{j} p^j p^{N-j}$$

This is a one-tailed test that is based on the assumption that the observed total sum of perfect matches is greater than the expected total sum.

As was indicated above, this test is exact. However, it can be conservative if the parameter p is estimated from the data in terms of relative frequencies. In this case, it can occur that the estimate of p reflects data characteristics that deviate from the characteristics of the population, and the test becomes conservative or biased.
The variance of the binomial distribution is $S^2(P) = Npq$ with $q = 1 - p$. The mean is $E(P) = Np$. Using these terms, one can approximate the normal distribution when N is large enough. One obtains the test statistic

$$Z_{Bin} = \frac{\sum_i m_{ii} - Np}{\sqrt{Npq}}$$

Z_{Bin} is equivalent to P for large N. It is also interesting to note that Z_{bin} is equivalent to Z for large numbers of rating categories. This is shown in the appendix.

4.3 A Note on the Covariance Structure resulting from the Multinomial Distribution

In rater agreement, the usual sampling scheme is multinomial. That is, only the total number of responses is fixed. From this sampling scheme, a certain covariance structure between the cells results (see, e.g., Agresti, pp. 579 - 580 [14]). This issue is important because the test statistics discussed in this article require independence of the main diagonal cells in the table.
Let the observation k of the N cross-classified data in the $L = (I \times I)$-table be denoted by $\boldsymbol{Y}_k = (Y_{k1}, ..., Y_{kL})$, where $Y_{kl} = 1$ if the observation k falls in cell l, and $Y_{kl} = 0$ otherwise. Since each observation falls in just one cell, it follows that

$$\sum_{l=1}^{L} Y_{kl} = 1,$$

and

$$Y_{kl} Y_{km} = 0,$$

if $l \neq m$ (Remark: m is used here as an index variable for the cells). As a consequence, the expected values are

$$E(Y_{kl}) = P(Y_{kl} = 1) = \pi_l = E(Y_{kl}^2),$$

and

$$E(Y_{kl} Y_{km}) = 0$$

if $l \neq m$. In matrix notation: $E(\boldsymbol{Y}_k) = \boldsymbol{\pi}$ and $cov(\boldsymbol{Y}_k) = \boldsymbol{\Sigma}$, $k = 1, \ldots, N$. What is left at this point is the determination of the variance-covariance matrix $\boldsymbol{\Sigma}$. From the preceding results it follows that the elements on the main diagonal are

$$\sigma_{ll} = var(Y_{kl}) = E(Y_{kl}^2) - E(Y_{kl})^2 = \pi_l(1 - \pi_l).$$

The covariances on the off-diagonal $l \neq m$ are given by

$$\sigma_{kl} = cov(Y_{kl}, Y_{km}) = E(Y_{kl} Y_{km}) - E(Y_{kl})E(Y_{km}) = -\pi_l \pi_m.$$

Finally, the matrix Σ has the form

$$\Sigma = diag(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}'.$$

These results suggest that the frequencies cells in the rater contingency table are not independent of each other. Therefore, the frequencies m_{ii} in the main diagonal ($i = 1, \ldots, I$) are correlated. Both test statistics require independent components: In the binomial test, the m_{ii} terms have to be independent. In Stouffer's approach, the independence condition applies to the z_{ii} terms, which in turn are determined by the m_{ii}. As a consequence, the required assumptions for using the test do not completely hold. Therefore, in the next section, simulation studies are performed to show the influence of this covariance structure on the performance of the tests.

5 Simulation Studies

In this section, we present a simulation study of Stouffer's Z and the binomial test. The objective of the simulation study is to investigate *type I* and *type II error* behavior of the proposed test statistics for common rating situations. For this purpose, contingency tables varying in size, agreement structure, and sample size had to be generated. The number of simulation runs performed under each combinatorial setting of the parameters is fixed to N_{Sim}=2000. In addition to Stouffer's Z and the binomial test, two Likelihood Ratio tests *(LR)* for the well known *quasi independence model (QI)*,

$$\log(m_{ij}) = \lambda + \lambda_i^A + \lambda_j^B + \delta_i I(i = j)$$

which is commonly used in the context of agreement modeling (see for example [4] or [14]) are given. $I(i = j)$ is an indicator function with

$$I(i = j) = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

The λ- parameters specify intercept and main effects, whereas δ_i permit the diagonal agreement cells of the contingency tables to deviate from independence. The null hypothesis under the first LR test assumes, that all δ_i parameters equal zero. Thus, this $LR^{(1)}$ test statistic is determined by the difference in deviances between the more parsimonious log-linear main effect model (ME), $\log(m_{ij}) = \lambda + \lambda_i^A + \lambda_j^B$, and the more complex QI model. As usual, $LR^{(1)}$ is χ^2- distributed with $df = df_{ME} - df_{QI}$. A second $LR^{(2)}$ test is performed which compares the intercept-only model $\log(m_{ij}) = \lambda$ with QI. The corresponding *Type I error* rates are studied in the following section.

5.1 Type I Error Behavior

Analyzing *type I error* behavior of the proposed test statistics is carried out by determining the failure rates of accepting H_0 (H_0 stating random agreement) if given data really stems from random agreement.
To test *type I error* rates and to obtain rejection rates for the statistics of interest, numerous random-agreement contingency tables must be created by simulation. One way to obtain these contingency tables is to take draws from a multinomial distribution with fixed homogeneous probability structure. Performing the proposed tests ($LR^{(1)}$, $LR^{(2)}$, Z, Z_{Bin}) on the artificially generated random-agreement data yields the following results for different sizes I of the contingency tables and different sample sizes, N:

N	$\boldsymbol{LR^{(1)}}$	$\boldsymbol{LR^{(2)}}$	$\boldsymbol{Z}$	$\boldsymbol{Z_{Bin}}$	$\boldsymbol{Z_{Bin}}$ $\mathbf{H_0}$, $\boldsymbol{Z}$ $\mathbf{H_0}$	$\boldsymbol{Z_{Bin}}$ $\mathbf{H_0}$, $\boldsymbol{Z}$ $\mathbf{H_1}$	$\boldsymbol{Z_{Bin}}$ $\mathbf{H_1}$, $\boldsymbol{Z}$ $\mathbf{H_0}$	$\boldsymbol{Z_{Bin}}$ $\mathbf{H_1}$, $\boldsymbol{Z}$ $\mathbf{H_1}$
100	0.0620	0.0565	0.0275	0.0405	0.9595	0.000	0.0130	0.0275
300	0.0410	0.0470	0.0200	0.0485	0.9515	0.000	0.0285	0.0200
500	0.0540	0.0475	0.0290	0.0495	0.9505	0.000	0.0205	0.0290
1.000	0.0550	0.0605	0.0215	0.0520	0.9480	0.000	0.0305	0.0215
5.000	0.0515	0.0465	0.0265	0.0555	0.9445	0.000	0.0290	0.0265
10.000	0.0590	0.0565	0.0205	0.0450	0.9550	0.000	0.0245	0.0205
50.000	0.0480	0.0525	0.0235	0.0500	0.9500	0.000	0.0265	0.0235
500.000	0.0445	0.0440	0.0270	0.0555	0.9445	0.000	0.0285	0.0270

Table 2: Type I errors for test statistics, and cross classification rates - table size I = 3

The values in Tables 5.1 to 5.1 suggest the following interpretation: For almost all table sizes, I, and sample sizes, N, the statistics Z, Z_{Bin}, $LR^{(1)}$ and $LR^{(2)}$ hold the α- level of .05. As can be seen, Z is conservative compared to the other test statistics as its rejection rate is always lower than the corresponding rate for Z_{Bin} (and the LR tests also). This can also bee seen in the four right-most columns of the tables, which give cross classification rates: There is no case in which the less rigorous test Z_{Bin} decides in favor of H_0 and Z decides in favor of H_1. Comparing rejection

N	$LR^{(1)}$	$LR^{(2)}$	Z	Z_{Bin}	Z_{Bin} H_0, Z H_0	Z_{Bin} H_0, Z H_1	Z_{Bin} H_1, Z H_0	Z_{Bin} H_1, Z H_1
100	0.0700	0.0570	0.0370	0.0595	0.9405	0.000	0.0225	0.0370
300	0.0560	0.0565	0.0410	0.0525	0.9475	0.000	0.0115	0.0410
500	0.0665	0.0630	0.0355	0.0515	0.9485	0.000	0.0160	0.0355
1.000	0.0600	0.0560	0.0330	0.0570	0.9430	0.000	0.0240	0.0330
5.000	0.0435	0.0465	0.0340	0.0505	0.9495	0.000	0.0165	0.0340
10.000	0.0545	0.0545	0.0395	0.0590	0.9410	0.000	0.0195	0.0395
50.000	0.0665	0.0665	0.0380	0.0560	0.9440	0.000	0.0180	0.0380
500.000	0.0535	0.0505	0.0300	0.0445	0.9555	0.000	0.0145	0.0300

Table 3: Type I errors for test statistics, and cross classification rates - table size I = 5

N	$LR^{(1)}$	$LR^{(2)}$	Z	Z_{Bin}	Z_{Bin} H_0, Z H_0	Z_{Bin} H_0, Z H_1	Z_{Bin} H_1, Z H_0	Z_{Bin} H_1, Z H_1
100	0.0760	0.0655	0.0455	0.0455	0.9545	0.000	0.0000	0.0455
300	0.0585	0.0545	0.0425	0.0590	0.9410	0.000	0.0165	0.0425
500	0.0580	0.0610	0.0340	0.0465	0.9535	0.000	0.0125	0.0340
1.000	0.0500	0.0520	0.0405	0.0465	0.9535	0.000	0.0060	0.0405
5.000	0.0490	0.0480	0.0280	0.0430	0.9570	0.000	0.0150	0.0280
10.000	0.0495	0.0475	0.0355	0.0450	0.9550	0.000	0.0095	0.0355
50.000	0.0575	0.0540	0.0400	0.0520	0.9480	0.000	0.0120	0.0400
500.000	0.0550	0.0500	0.0325	0.0465	0.9535	0.000	0.0140	0.0325

Table 4: Type I errors for test statistics, and cross classification rates – table size I = 7

N	$LR^{(1)}$	$LR^{(2)}$	Z	Z_{Bin}	Z_{Bin} H_0, Z H_0	Z_{Bin} H_0, Z H_1	Z_{Bin} H_1, Z H_0	Z_{Bin} H_1, Z H_1
100	0.0655	0.0775	0.0495	0.0495	0.9505	0.000	0.0000	0.0495
300	0.0685	0.0625	0.0435	0.0435	0.9565	0.000	0.0000	0.0435
500	0.0630	0.0560	0.0440	0.0440	0.9560	0.000	0.0000	0.0440
1.000	0.0505	0.0475	0.0380	0.0510	0.9490	0.000	0.0130	0.0380
5.000	0.0540	0.0535	0.0445	0.0560	0.9440	0.000	0.0115	0.0445
10.000	0.0435	0.0460	0.0350	0.0425	0.9575	0.000	0.0075	0.0350
50.000	0.0485	0.0545	0.0405	0.0495	0.9505	0.000	0.0090	0.0405
500.000	0.0425	0.0470	0.0415	0.0565	0.9435	0.000	0.0150	0.0415

Table 5: Type I errors for test statistics, and cross classification rates – Table Size I = 9

rates of Z and Z_{Bin} over different table sizes, reveals a convergence tendency of both statistics. Another simulation which covers table sizes up to I = 40 yields the results shown in Figure 5. In the appendix, a proof is given that Z_{Bin} and Z coincide for $I \to \infty$.

5.2 Type II Error Behavior

Analyses of *type II error* rates is usually performed by systematically generating data with a priori agreement structure, and determining the rate of wrong acceptance of H_0 (H_0 again assumes random agreement). However, in this study we use the complementary measure, the rate of correct rejection of H_0, which corresponds to the power of the test statistics. Creating random contingency tables which exhibit predefined degrees of agreement structure is somewhat more difficult. The creation of agreement probability structures in this section is inspired by Brennan and Prediger's κ_n. In *type II error* simulations, the table size I, the sample size N, and the agreement structure are varied systematically. While sample size and table size need no further explanation, the term "agreement structure" or, in short, "agreement", needs concrete specification. One possibility to control for rating agreement is to adapt the sum of probabilities of the cells in which the two decision making units agree exactly. Its minimum value, the case of random agreement, is $1/I$;

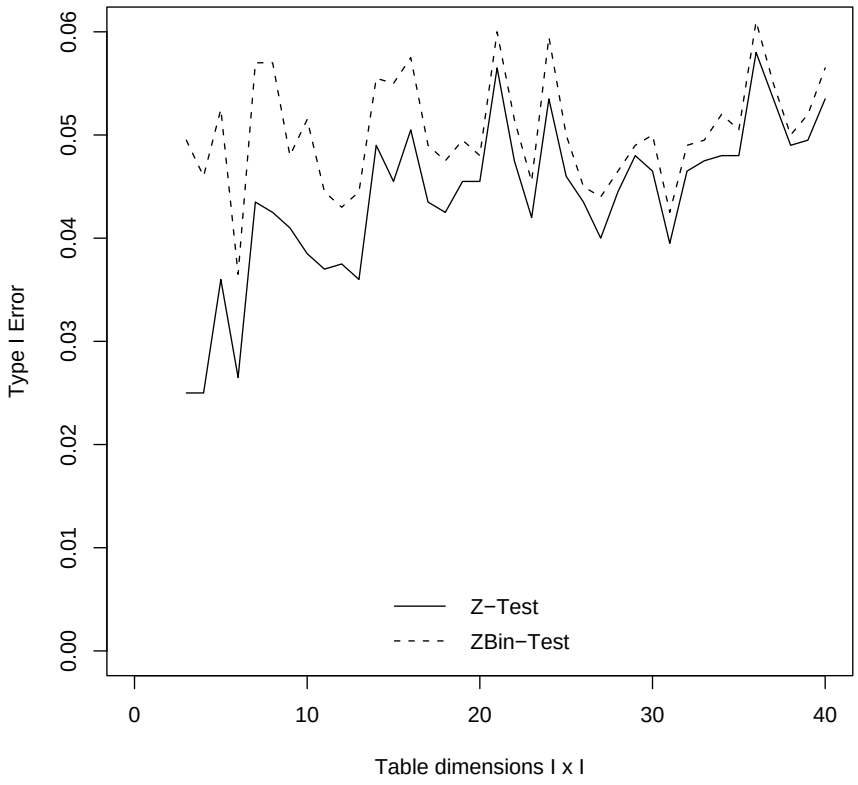


Figure 5: Type I error rates for Z (solid line) and Z_{bin} (hashed line) for tables of increasing size

its maximum value is one. In the simulation runs, the variable *AGRATE,* varied in equal intervals between zero and one, is used to indicate degree of agreement. The sum of the agreement diagonal probabilities p_{ag} is given by:

$$p_{ag} = 1/I + (1 - 1/I) * AGRATE$$

Increasing $AGRATE$ in equal steps yields corresponding intervals for p_{ag}. Given a specific interval, a uniform random number is drawn. This number represents the sum of the probabilities in the agreement diagonal. Specific cell probabilities for the agreement diagonal are then generated by uniformly distributing the probability mass of the agreement diagonal over its cells. The same procedure is carried out for the off-diagonal, that is, the disagreement cells. Thus, for each combination of the controlled variables table size I, agreement interval of *AGRATE,* and sample size N, the probability structures, and finally, samples are generated by drawing from a multinomial distribution.
In this simulation study, $AGRATE$ varies in 35 steps, 4 table sizes were chosen (3 x 3, 5 x 5, 7 x 7 and 9 x 9), and the sample size ranged from 100 up to 500.000, in unequal, increasing intervals. The results of this simulation procedure are summarized in Figures 6 to 9.

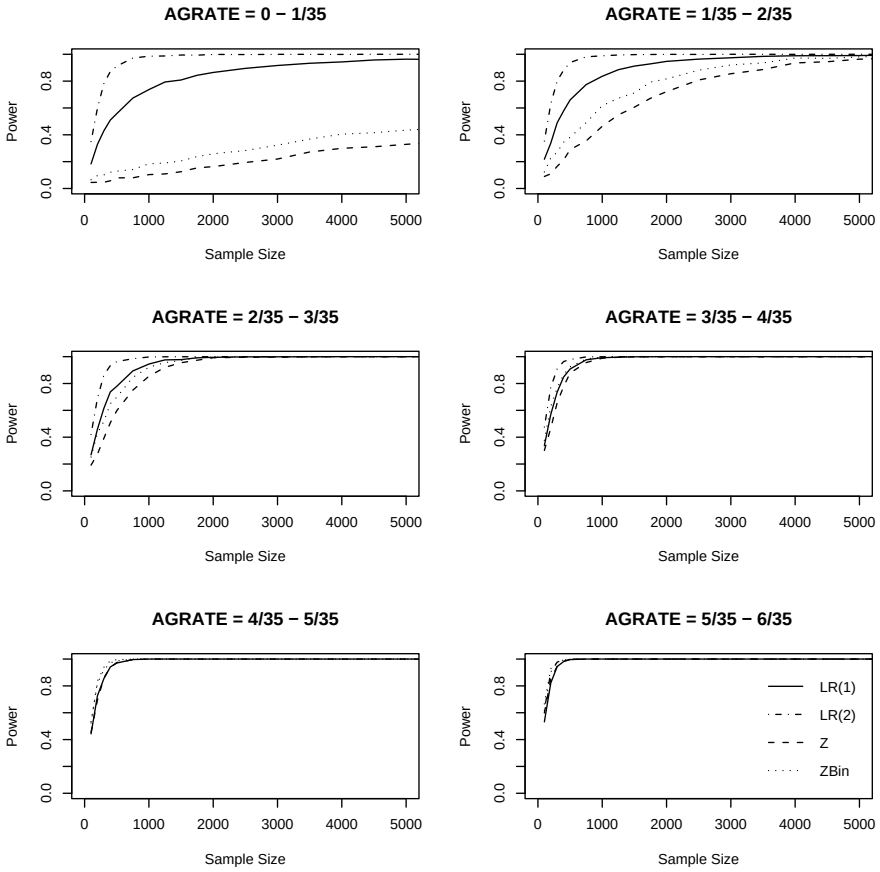


Figure 6: Simulation results for power of LR, Z and Z_{Bin} tests, table size $I = 3$

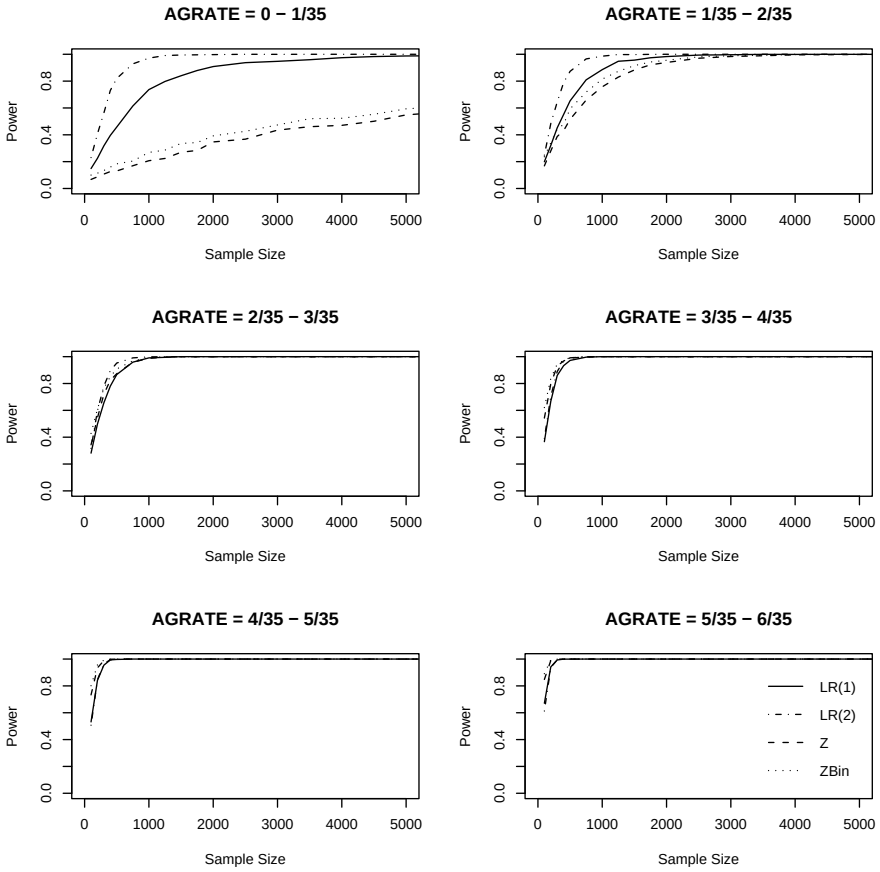


Figure 7: Simulation results for power of LR, Z and Z_{Bin} tests, table size $I = 5$

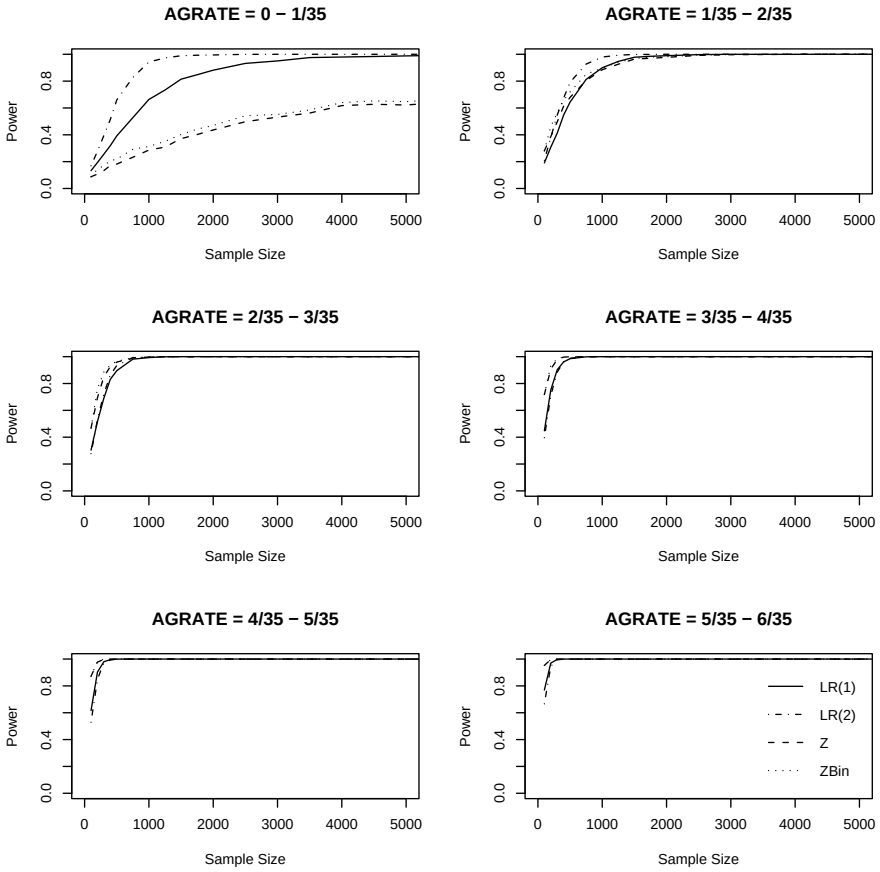


Figure 8: Simulation results for power of LR, Z and Z_{Bin} tests, table size $I = 7$

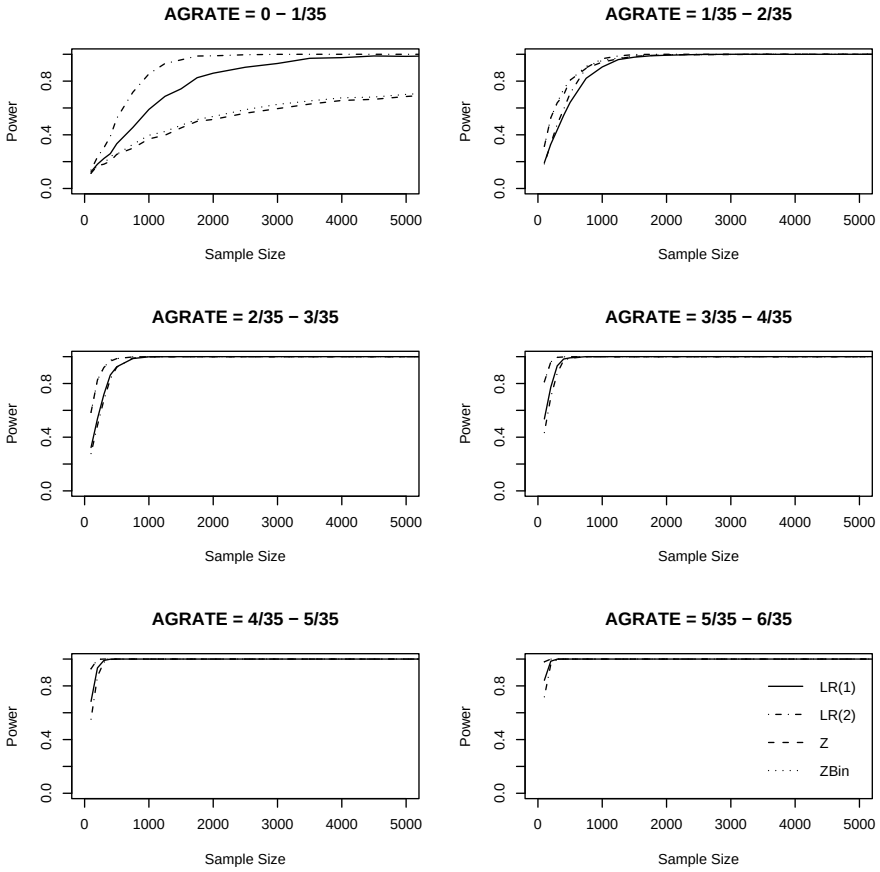


Figure 9: Simulation results for power of LR, Z and Z_{Bin} tests, table size $I = 9$

All simulation runs, including the ones with table sizes between 3 x 3 and 9 x 9, show similar patterns: If the real agreement effect in the data is extremely small, e.g. $AGRATE \in [0;1/35]$, LR tests provide considerably higher power than Z_{Bin} or Z. Independent of table size I, Z_{Bin} and Z show a severe lack of power in this case. The power of these measures increases only very slowly as sample size increases. However, as soon as the effect of agreement becomes stronger, Z and Z_{Bin} catch up very quickly, coincide and sometimes even outperform LR tests in terms of power. When size I of the table is increased, the power of Z and Z_{Bin} is approximately the same. An application example using a real data set is given in the next session.

6 Data Example

The following example uses the data from Table 2. We calculate the expected cell frequencies for $\hat{ra}$ and κ_n as 129/9 = 14.33. As was reported above, raw agreement of the two psychiatrists who provided the data in Table 2 is $\hat{ra} = 0.74$. The standardized deviates, z_{ij}, that result for the null model that is used for Brennan and Prediger's κ_n and for the significance test of the coefficient of raw agreement, are displayed in 6. The estimated expected frequencies for this model are $\hat{m}_{ij} = \frac{N}{I^2} = \frac{129}{3^2} = 14.33$.

		Psychiatrist 2: Severity of Depression		
		1	2	3
Psychiatrist 1:	1	-0.88	-3.52	-3.79
Severity of Depression	2	-3.26	-2.99	-1.67
	3	1.23	-2.99	17.87

Table 6: Standardized deviates for the null model for the data in table 1

Table 6 shows that the high rate of exact raw agreement is carried mostly by Cell 3 3. This cell contains far more judgements than expected based on the null model. The other two diagonal cells make negative contributions by containing fewer cases than expected.
We now employ first the probability pooler, *Stouffer's Z*, and then the binomial test to determine whether 74% agreement is better than chance agreement as determined by the null model. We calculate Z = (-0.88 - 2.994 + 17.873)/$\sqrt{3}$ = 8.082. This value is larger than the critical $z_{0.05}$ = 1.96, and we conclude that the portion of raw agreement is significantly better than $1/I = 0.333$. For the binomial test, we estimate p as $1/I = 0.333$, and $q = 0.667$. Inserting yields the z-approximation Z_{Bin} = (96 - 129(.333))/$\sqrt{(129 \cdot 0.333 \cdot 0.667)}$ = 53/5.35 = 9.90. This value also suggests rejecting the null hypothesis of agreement that is no better than 0.333.

7 Summary and Discussion

In this article, we showed that significance tests can be devised for the coefficient of raw agreement which, thus far, has only been used as a descriptive measure. The significance test is based on the fact that Brennan and Prediger's κ_n and ra are linearly dependent upon each other. Therefore, the same base model can be employed to estimate expected cell frequencies. This base model is the log-linear null model. The resulting expected cell frequencies serve as reference values, and can be used to devise significance tests. Stouffer's Z and the binomial test are proposed as significance tests. Thus, the coefficient of raw agreement can be used for both, descriptive and test purposes. Simulation studies showed that both test statistics perform well.
The comparison of Figures 6 to 9 shows an interesting characteristic of the present results. For all tables and weak effects (AGRATE 0 - 1/35), the power of the two tests discussed here fails to exceed 0.8, even when the sample size is as large as 5000. Even for effects of midrange magnitude,

the tests approach $p = 1$ only slowly. Given weak effects, in larger tables, the power of the proposed tests grows much faster than in smaller tables. This pattern is worth discussing because, for any given sample size, the cell size in smaller tables is larger than in larger tables. The reason for the lower power is the covariance structure discussed above. When tables are small, cells are more dependent than in larger tables. In 3 x 3 tables, a third of the cells is located in the main diagonal. In 9 x 9 tables, only 11% of the cells are in the main diagonal. The reduction in dependency is reflected in the present results. Even when the cell size is smaller (larger tables; given sample size), the tests display more power. Still, the two tests discussed here perform rather well even for small tables.

One may wonder whether the two LR tests that were also used in the present simulations could be recommended as significance tests for the measure of raw agreement. Two arguments prevent one from making this recommendation. First, the $LR^{(1)}$ test used here uses a different base model than appropriate for the measure of raw agreement. It uses the main effect model instead of the null model. Second, and this argument applies to both LR tests, the tests are sensitive to deviations from the base model that do not reflect rater agreement. It is easy to show that the $LR^{(2)}$ test can indicate strong deviations from its base model even if there is zero rater agreement. In fact, when $ra = 0$, the $LR^{(2)}$ test will signal deviations.

Because of this characteristic of the LR tests, the simulation results suggest superior power of the LR tests over the Z and the binomial tests that may not exist. The Z and the binomial tests are sensitive only to deviations caused by agreement beyond the degree proposed in the base model. The LR test are also sensitive to deviations in other parts of the table.

The approach to a significance test for the coefficient of raw agreement proposed in this article can be generalized in several ways. For example, raw agreement among three or more raters can be studied. In addition, more complex base models can be considered, for example base models with covariates.

The advantages of the coefficient of raw agreement over Cohen's κ include that ra is not marginal-dependent. It shares this characteristic with Brennan and Prediger's κ_n [3] and Martín and Femia's delta [15]. ra approximates the maximum value of 1 even if the marginal distribution is very uneven. Now, that a significance test for ra exists, ra can be considered as a measure of rater agreement that does not come with the interpretational problems of κ.

References

[1] von Eye A, Mun EY. *Modeling rater agreement - Manifest variable approaches*. Lawrence Erlbaum: Mahwah, NJ, 2005.

[2] Cohen J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 1960; **20**: 37 - 46.

[3] Brennan RL, Prediger DJ. Coefficient kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement* 1981; **41**: 687 - 699.

[4] Tanner MA, Young MA. Modeling agreement among raters. *Journal of the American Statistical Association* 1985; **80**: 175 - 180.

[5] Uebersax JS, Grove WM. Latent class analysis of diagnostic agreement. *Statistics in Medicine* 1990; **9**: 559 - 572.

[6] Schuster CA. A mixture model approach to indexing rater agreement. *British Journal of Mathematical and Statistical Psychology* 2002; **55**: 289 - 303.

[7] von Eye A, Schuster C. Log-linear models for rater agreement. *Multiciência* 2000; **4**: 38 - 56.

[8] Fleiss JL. *Statistical methods for rates and proportions*, 2nd ed., Wiley: New York, 1981.

[9] Hsu LM, Field R. Interrater agreement measures: Comments on Kappa$_n$, Cohen's Kappa, Scott's π, and Aickin's α. *Understanding Statistics* 2003; **2**: 205 - 219.

[10] von Eye A, Sörensen S. Models of chance when measuring interrater agreement with kappa. *Biometrical Journal* 1991; **33**: 781-787.

[11] Stouffer SA, Suchman EA, DeVinney LC, Star SA, Williams RM Jr. *The American soldier: Adjustment during Army life* (vol. 1). Princeton University Press: Princeton, NJ, 1949

[12] Darlington RB, Hayes AF. Combining independent p-values: Extensions of the Stouffer and binomial models. *Psychological Methods* 2000; **5**: 496 - 515.

[13] Strube MJ. Combining and comparing significance levels from nonindependent hypothesis tests. *Psychological Bulletin* 1985; **97**: 334 - 341.

[14] Agresti A. *Categorical data analysis*, 2$^{\text{nd}}$ ed., Wiley: New York, 2002

[15] Martín AA, Femia MP. Delta: A new measure of agreement between two raters. *British Journal of Mathematical and Statistical Psychology* 2004; **57**: 1 - 19

A Appendix

The equivalence between Stouffer's Z and the normal approximation Z_{bin} of the binomial test for a large number of rating categories I

As is well known, the binomial distribution can be approximated by the a normal distribution when N is large. In this appendix, we show that Z_{bin} and Z are equivalent if I is large. The idea is to transform Z into Z_{bin}. We use the statistics

$$Z_{bin} \;=\; \frac{\sum_i m_{ii} \;-\; Np}{\sqrt{Np(1 \;-\; p)}}$$

and

$$Z \;=\; \frac{1}{\sqrt{I}} \sum_i \frac{m_{ii} \;-\; N \;/\; I^2}{\sqrt{N \;/\; I^2}}$$

if z_{ii} is inserted in Z. For the binomial test, the parameter p is estimated by $1/I$. By applying this assumption for the calculation of Stouffer's Z, one obtains

$$Z \;=\; \sqrt{p} \;\left[\frac{1}{\sqrt{Np^2}} \;\left(\sum_i m_{ii} \;-\; Np \right) \right]$$

Note that the last term Np follows from $\sum_i \frac{N}{I^2} \;=\; \frac{I \;N}{I^2} \;=\; Np$.

For the denominator in the equation for Z, we obtain $\sqrt{Np^2} \;=\; p\sqrt{N}$, and, therefore,

$$Z \;=\; \frac{\sqrt{p}}{p\sqrt{N}} \;\left(\sum_i m_{ii} \;-\; Np \right) \;=\; \frac{\sum_i m_{ii} \;-\; Np}{\sqrt{Np}}$$

This expression is similar to Z_{bin} with the exception that, in the denominator, in Z_{bin}, we find $\sqrt{Np \;(1 \;-\; p)} \;=\; \sqrt{Np \;-\; Np^2}$ compared to $\sqrt{Np}$ in the Stouffer formula. Now, if $I \;\rightarrow\; \infty$, it follows that the binomial parameter $p \;\rightarrow\; 0$. This implies that the denominators of both test statistics are asymptotically equivalent. Since the difference between these two test statistics concerns only the denominator, it follows that $Z \;\rightarrow\; Z_{bin}$ for $I \;\rightarrow\; \infty$. This result also shows that, for small numbers of rating categories, the two test statistics produce different outcomes. In the above example, the number of rater categories is 3, and the difference in the test statistics is about 1.8. The difference between the two statistics may not always be negligible (as it is in the above example), in particular when the significance tests suggest contradictory statistical decisions.