Psychology Science, Volume 47, 2005 (3/4), p. 401-423

Combinatoric search for types and antitypes

STEFAN VON WEBER[1], ALEXANDER VON EYE[2] & ERWIN LAUTSCH[3]

Abstract

The combinatoric strategy of searching for types, proposed by Kieser and Victor (1991), is embedded into the frame of standard CFA methodology. As part of this approach, additional, new terms are included in the search procedure. Including these terms leads to clear improvements. A comparison of the new, combinatoric method with four other CFA test procedures shows the power of the new algorithm. It performs better than the comparison methods under all simulated conditions.

Key words: contingency table, configural frequency analysis, CFA, search for types, statistical test, simulation, continuity correction

[1] Prof. Dr. Stefan von Weber, Fachhochschule Furtwangen, FB Maschinenbau und Verfahrenstechnik, PSF 3840, D-78027 Villingen-Schwenningen; email: webers@hs-furtwangen.de

[2] Prof. Dr. Alexander von Eye, Michigan State University, Department of Psychology, 107D Psychology Building, East Lansing, MI 48824-1116, USA; email: voneye@msu.edu

[3] Prof. Dr. Dr. Erwin Lautsch, Universität Kassel, FB5: Gesellschaftswissenschaften, Nora-Platiel-Straße 1, D-34127 Kassel; email: erla@uni-kassel.de

1. Introduction

In 1991, Kieser and Victor published "a test procedure for an alternative approach to Configural Frequency Analysis." This procedure is based on Victor's (1989) alternative statistical type concept. In this approach, types or antitypes are constituted by cells that possess an additional probability mass (in the case of types) or lack it (in the case of antitypes), compared to the expectation that is based on some base model which reflects the relation of the attributes of the remaining cells. In their 1991 article, Kieser and Victor considered the case of a single type or antitype, that is, the case of a single structural zero. The expected frequency for cell k, $\hat{e}_k$, was estimated using Goodman's (1968) ML estimator (cf. Victor, 1983). In the section on generalizations, Kieser and Victor formulated a general procedure for the exploratory search for types in a given contingency table, C with I cells. The authors write (1991, pp. 91 - 92) "

1. Fix a number, m, $1 \leq m \leq I$, as the number of types to search for
2. Search a minimal set $U \subset I$, $|U| \leq m$, so that $C(I \backslash U)$ is quasi-independent and $C(U)$ is a set of type configurations."

In their 1999 article, Kieser and Victor applied this concept in the context of log-linear modeling. The use of log-linear models has the advantage that new models and, thus, CFA base models, can be created with simple changes in the design matrix. Types are found as large residuals with a positive deviation, and antitypes are large negative residuals. A multiplicity issue arises from the fact that for types or antitypes to be constituted, both the null hypotheses for the non-type/non-antitype cells and the alternative hypotheses for the type/antitype cells must be retained.

2. The pure combinatoric search algorithm

Let a contingency table $\{n_{ijk}\}$ be given with N_c cells (configurations, symptom configurations) where n_{ijk} is the observed frequency of cases with symptom configuration (i, j, k). The dimension of the table is $d \geq 2$, with I rows, J columns, and so forth. With Kieser and Victor, we assume that whereas most of the cells obey the base model of independence, only a few type cells or antitype cells are outliers. In the following paragraphs, we describe the combinatoric search algorithm.

Step 1: Estimate the maximum number $n_{t,\,max}$ of types to search for. In this article, we used the formula $n_{t,\,max} = \text{round}(df^{0.5} - 0.49)$, where df is the degree of freedom of the table. For example, if df has a value of $4 \leq df \leq 8$, we find $n_{t,\,max} = 2$. Then, mark a first cell of the table, and set number n_t of actually assumed type cells to $n_t = 1$.

Step 2: Calculate the so-called Victor expectancies, $\hat{e}^*_{ijk}$, of all cells using the Deming-Stephan EM algorithm. Marked cells are assumed to be type cells or antitype cells. Then, calculate two X^2 sums, the first from all marked cells, and the second from all unmarked cells:

$$X_t^2 = \sum_{marked} \frac{(n_{ijk} - \hat{e}_{ijk}^*)^2}{\hat{e}_{ijk}^*}$$

$$X_{non}^2 = \sum_{unmarked} \frac{(n_{ijk} - \hat{e}_{ijk}^*)^2}{\hat{e}_{ijk}^*}$$

Using X_t^2 and X_{non}^2, calculate the F-statistic

$$F = \frac{X_t^2 (N_c - n_t)}{X_{non}^2 n_t}$$

Step 3: Mark the next cell and repeat Step 2. Continue until all cells are tested. If $n_{t,\,max} > 1$ holds, continue with pairwise combinations of marked cells (case $n_t = 2$), e.g., (1, 2), (1, 3), ... If $n_{t,\,max} > 2$ holds, continue with triples of marked cells (case $n_t = 3$), e.g., (1, 2, 3), (1, 2, 4), ... This combinatoric search continues until $n_t = n_{t,\,max}$ holds. From all calculated F-values, we take the maximum value and store the vector of Victor expectancies $\hat{e}_{ijk}^*$, but not the pattern of type cells.

Step 4: The confirmatory hypotheses tests are performed using the local test of Dunkl and von Eye (1990) with the Victor expectancies $\hat{e}_{ijk}^*$ that result in Step 3,

$$X = \frac{n_{ijk} - \hat{e}_{ijk}^*}{\sqrt{\sigma_{ijk}^{*2}\,(1 - K)}},$$

with variance $\sigma_{ijk}^{*2} = \hat{e}_{ijk}^*(\hat{e}_{ijk}^* + 0.5)/(\hat{e}_{ijk}^* - 0.5)$. Here, K is the table-specific continuity correction introduced by von Weber, Lautsch, and von Eye (2003a; cf. von Weber, Lautsch, & von Eye, 2003b). The significance threshold α can be adjusted using, for example, Holm's procedure.

3. Improvements of the F statistic

In the simulations, the size of the tables was varied, that is, we used tables that differed in dimensionality, degrees of freedom, and mean frequencies (more detail follows below). The simulations showed that the simple F-statistic described above can be improved. The first improvement is based on the observation that type cells often are constituted by cells with above average frequencies. Antitypes often are constituted by cells with below average frequencies. We formulated three different terms, called frequency bonuses, B_1, B_2, and B_3 to evaluate this observation,

$$B_1 = | n_{ijk} - m | /m,$$
$$B_2 = (n_{ijk} - m)^2/m,$$
$$B_3 = \sqrt{n_{ijk} - m}/m.$$

Each summand of X_t^2 was multiplied by either B_1, B_2, or B_3. Simulating tests for tables of dimensions $d = 2$ and $df = 8$ resulted in β-errors as shown in Figure 1. The β-errors for tables with dimension $d = 3$ and $df = 4$ are shown in Figure 2.

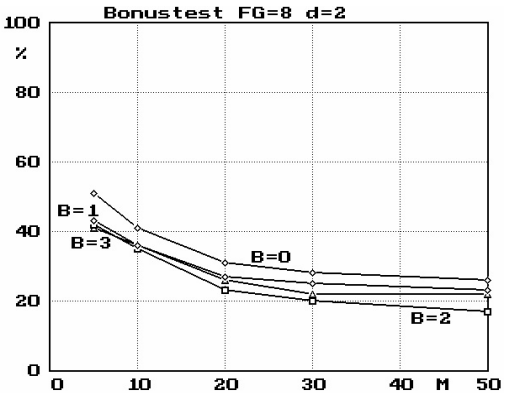

Figure 1:
β-errors of types for different frequency bonuses over mean cell frequency m,
for $d = 2$ and $FG = df = 8$

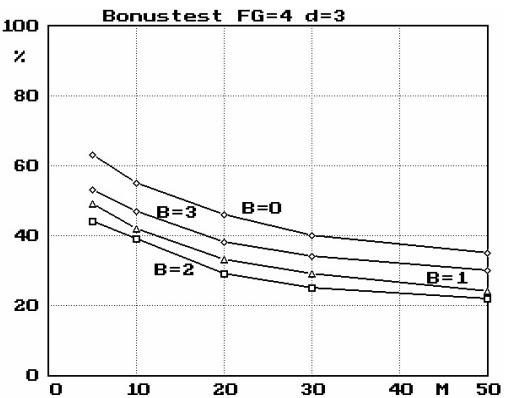

Figure 2:
β-errors of types for different frequency bonuses over mean cell frequency m,
for $d = 3$ and $FG = df = 4$

Both figures suggest that the β-errors are minimal for B_2. Therefore, B_2 was selected to become part of the search statistic F.

Detecting antitypes is much more difficult than detecting types. We use the type strength τ (see below and von Weber, 2000, von Weber et al., in press) to declare the lack of probability density. A type strength of τ = 2 multiplies the base frequency e^*_{ijk} by (1 + τ) = 3, and, for antitypes, divides the base frequency by (1 + τ). Consider, for example, a base frequency of cell *ijk* of e^*_{ijk} = 20. A type cell is then expected to contain e_{ijk} = 20·3 = 60 cases. An antitype cell is then expected to contain e_{ijk} = 20/3 = 7 cases. Calculating the X^2-component for the type cell, one obtains $X^2 = \frac{(60 - 20)^2}{20} = 80$. For the antitype cell, one gets $X^2 = \frac{(7 - 20)^2}{20} = 8.5$. Obviously, the X^2-component is much larger for type cells. We performed simulations in which we varied bonuses that would make the search for antitypes easier. The resulting optimal bonus is

$$A = \frac{\hat{e}^*_{ijk}}{n_{ijk}}.$$

(If n_{ijk} < 3, we use A = $\hat{e}^*_{ijk}$ /3.) The antitype bonus follows from the relation $\hat{e}^*_{ijk} = n_{ijk}(1+\tau)$, for antitypes. To obtain the same X^2 component as in the type case, one would have to use the factor $A^2 = (1 + \tau)^2$. However, our simulations suggest that bonus A leads to better results than bonus A^2. Bonus A is multiplied with each summand of X^2_i, if for the cell marked as possible antitype $n_{ijk} < \hat{e}_{ijk}$ holds. Here, $\hat{e}_{ijk}$ is the common expectancy of independence. The following figures 3 and 4 illustrate the effects of the bonuses A and B.

The figures suggest that

- Antitypes are detectable only if the mean cell frequency is $m > 20$
- Types become harder to detect if an antitype bonus with $m < 20$ is used.

Therefore, our algorithm uses the antitype bonus A in the calculation of the F-statistics only for tables with $m \geq 20$.

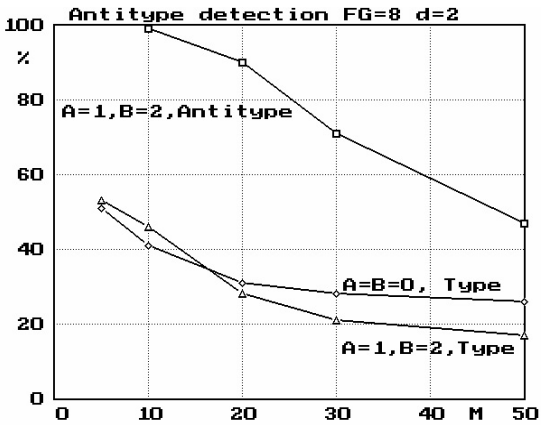

Figure 3:
β-errors of types and antitypes for different frequency-bonuses over mean cell frequency *m*, for FG = *df* = 8 and *d* = 2

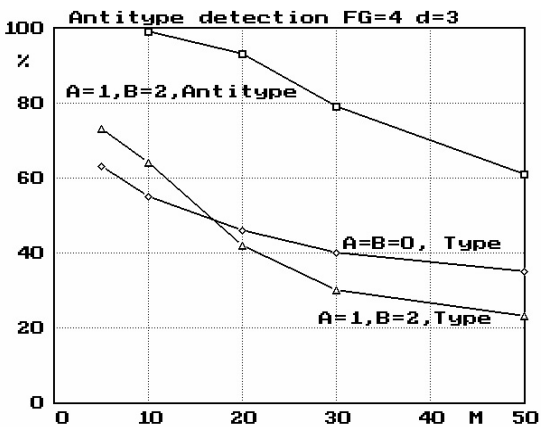

Figure 4:
β-errors of types and antitypes for different frequency-bonuses over mean cell frequency *m*, for *FG* = *df* = 4 and *d* = 3

3. Test data generation for simulations

Data were generated with the same characteristics as in von Weber et al. (2003a). The resulting tables varied in the number of variables, *d*, that span the table, the degrees of freedom, *df*, and the sample size, *N*, specifically, the average number of cases per cell, $m = N/N_c$ with N_c indicating the number of cells. In addition, type strength, τ, and distribution type, DT, were varied as in the earlier studies.

The *type strength*, τ, describes the weight of a type or an antitype. τ is the main determinant for the estimation of the magnitude *K* of the continuity correction. The concept of *type strength* can be derived from Lienert's (1969) definition of a *contingency type*: a cell that constitutes a type contains more cases than expected based on the assumption of variable independence. In more general terms, types can be defined as constituted by cells in which the discrepancy between the observed relative frequency and expected probability that was estimated using some base model is large, given a particular definition of deviation from independence (Goodman, 1991;von Eye, Spiel, & Rovine, 1995).

Let n_{ijk} be the observed cell frequency and $\hat{e}^*_{ijk}$ the expected frequency, estimated using Victor's log-linear models of quasi-independence (Kieser & Victor, 1999; Victor, 1989; cf. Lautsch & von Weber, 2003). Then, the ratio $n_{ijk} / \hat{e}^*_{ijk}$ is an estimate of τ + 1 (see also the discussion of *relative risk* in von Eye & Gutiérrez Peña, in this issue). The difference n_{ijk} - $\hat{e}^*_{ijk}$ represents the surplus frequency in Cell (*i*, *j*, *k*) that constitutes a type. A cell with $n_{ijk} = \hat{e}^*_{ijk}$ thus carries a type strength of zero, and τ = 0. If n_{ijk} is twice as large as $\hat{e}^*_{ijk}$, one obtains $\tau = (n_{ijk} - \hat{e}^*_{ijk})/\hat{e}^*_{ijk}$, etc. A cell constitutes an antitype if $n_{ijk} < \hat{e}^*_{ijk}$. The strength τ of an antitype is defined by $\tau = (\hat{e}^*_{ijk} - n_{ijk})/n_{ijk}$. If the strength of an antitype is τ = 1, the observed frequency of Cell (*i*, *j*, *k*) is half the size of the expected frequency.

There is an obvious asymmetry in the formulation of types and antitypes, as was noted already in earlier publications (Lautsch & von Weber, 2003; von Weber et al., 2003a). The reason for this asymmetry is that a type strength of $\tau \to \infty$ should, in principle, be possible for antitypes. However, the logical lower limit for cell frequencies is zero, for two reasons. First, small cell probabilities can result in zero observed cases unless samples are very large, because observed frequencies manifest in natural numbers. Second, probabilities can be zero as in the case of structural zeros. In either case, there are natural limits to the specification that the observed frequency be half the size of the expected frequency if the strength of an antitype is τ = 1. This applies accordingly to greater strength values.

The maximum type strength, τ_{max}, of a contingency table is estimated by $\max_{ijk}(n_{ijk} - \hat{e}^*_{ijk})/m$, where *ijk* goes over all cells in the table, n_{ijk} is the observed cell frequency, $\hat{e}^*_{ijk}$ is the expected cell frequency, estimated under a suitable base model, and *m* is the average cell frequency. In simulations, tables with a priori determined type strength can be created. The above measures can then be used to estimate the value of τ_{max} for an observed table.

The distribution type, *DT*, indicates whether the cell-specific residuals are normally (*DT* = 0) or binomially (*DT* = 1 or *DT* = 2; see below) distributed. Another distribution type that can be considered is the hypergeometric distribution. The hypergeometric and the binomial

distribution approximate each other if the size of the population is much larger than the size of the sample.

The simulation. Data for the following simulations were generated in the following seven steps.

Step 1: Contingency tables with $d = 2$, 3, 4, and 5 were created. Of the resulting tables, only those were kept that had degrees of freedom within an a priori-specified interval. These intervals were [4, 5], [8, 10], [16, 20] for $d = 2$ and $d = 3$, [8, 10] and [16, 20] for $d = 4$, and $df = 26$ for $d = 5$.

Step 2: This step involved the random generation of the probabilities of the marginal distributions under the usual constraint that $0 < p < 1$ and $\Sigma p = 1$. Cell frequencies were then estimated under the assumption of variable independence, that is, for four variables, $p_{ijkl} = p_i\, p_j\, p_k\, p_l$. Thus far, the data conform exactly to the model of variable independence.

Step 3: This step involves determining the number of types and antitypes, N_t. The upper limit of this number is given by $1 \le N_t \le df$ and $1 \le N_t \le \text{round}(df^{1/2} - 0.49)$. The second of these conditions was chosen arbitrarily. However, it is conform to Kieser and Victor's (1999) concepts of CFA. The *df* of a table indicates the maximum number of independent hypotheses in a contingency table (Perli, 1985; Perli, Hommel, & Lehmacher, 1985).

Step 4. The position of each of the N_t types and antitypes was selected randomly. However, two constraints were placed. First, no marginal sum could contain summands adding up to more than 50% of its cases from type or antitype cells. Thus, for example, in 2 x 3 tables, no column can contain more than one type or more than one antitype. The second constraint was that type cells had to have an a priori frequency of $n_{ijk} > m$, and antitype cells had to have an a priori frequency of $n_{ijk} < m$.

The second constraint is new. In former papers it was not used for type cell or antitype cell generation. The reason is that observed tables (compare e.g. the tables in section 6) follow in a wide range this pattern, and the simulation results, especially the calculated β-values, will be closer to reality. For example, in Table 1 which presents Lienert's LSD data, the mean frequency is $m = 65/8 = 8.1$. The only type found by the combinatoric search has the observed frequency 20, the only antitype the frequency 0. It should be noted that this is not by necessity. Tables can be found in which below-average frequencies constitute types. Consider, for example, Table 2, the cancer data of Havemann et al. (1987). The mean frequency is $m = 1127/16 = 70.4$. The two types found have frequencies 64 and 703. Frequency 64 is somewhat smaller than *m*, i.e. the only exception from the rule in three tables. A third example can be found in Table 4, the crime data of Lautsch (2000). Here, the average cell size is $m = 1952/27 = 72.3$. The three types found have all frequencies greater than *m* (192, 105, 83), all seven antitypes have frequencies smaller than *m* (4, 5, 4, 40, 2, 45, 63).

Step 5. To obtain types and antitypes, the probabilities p_{ijkl} were now multiplied by factors of the form $(1 + \tau)$. Thereby, the type strength is an even random number between 1 and τ_{max}, where τ_{max} was set to 2 for the present simulations. Thus, only the values of $\tau = 1$ and $\tau = 2$ were used. For $\tau = 1$, the observed frequency is twice as large as the Victor-estimated

expected frequency. The ratio of thus randomly created types and antitypes was set to be 2:1. The first type was assigned the value of the maximum strength type, that is, $\tau_{max} = 2$. When there was more than one type, the following types were assigned linearly decreasing values, with the constraint that these values be greater or equal to one. For example, for a table with 4 types and antitypes, and a maximum type strength of $\tau_{max} = 2$, the four τ-values are 2, 1, 1, 1. The reason for the lower limit of $\tau = 1$ lies in the control of β. When a weight of τ is too small, the chance of reliably identifying a type and, even more so, an antitype, is very small.

Step 6. This step is needed to make sure that the condition $\Sigma p = 1$ holds. After adding a constant to each cell, this condition holds no longer. Therefore, a correction is needed. Consider the cell for which the a priori probability is $p = 0.055$. If, for this cell, the weight $\tau = 1$ is used, that is, if this cell is randomly selected to constitute a type, its probability changes to be $p^* = 0.055 \cdot 2 = 0.11$, and the sum of all cell probabilities increases from $\Sigma p = 1$ to $\Sigma p = 1.055$. Therefore, we reduce the probability for each cell proportionally by the factor 1/1.055. The type-cell then has a probability of $p^{**} = 0.11/1.055 = 0.1043$, and the sum of the thus corrected cell probabilities is $\Sigma p = 1$ again. This applies accordingly when more than one type or antitype is in a table.

Step 7. The last step of the simulation involves calculating the estimated expected cell frequencies. Each cell is multiplied by the a priori determined sample size, $N = N_c m$. The values of m used in this simulation were $m = 5$, 10, 20, 30, and 50. The estimated expected cell frequencies where thus $\hat{e}_{ijk} = Np_{ijk}$, where *ijk* goes over all cells in the table. A drawing error was then added to each cell, depending on the distribution type, *DT*.

The term *drawing error* is used here to denote the discrepancy between observed and expected frequencies that can be observed even under optimal sampling conditions. The distribution of this error depends on the selected model. The resulting values were rounded to be integers. Negative values were set to zero. It should be noted that this drawing error is the main source of errors in weakly frequented contingency tables. It prevents researchers from reliably identifying types and antitypes, in particular if they are of minor strength. Please note that the term *sampling error* denotes additional errors that may be hard to quantify. These errors reflect discrepancies between model and reality.

4. Tests and test procedures

The tests compared in this simulation are (1) the new *combinatoric search* (CS) proposed in this article, (2) Lautsch and von Weber's (2003) *new test procedure* (nPr), (3) Lienert's χ-component test (Chi), (4) Lehmacher's (1981) asymptotic hypergeomeric test with Küchenhoff's (1986) continuity correction (LK), and (5) the asymptotic test of Perli, Hommel, and Lehmacher (1984) (Pe). Each test was performed under the two-tailed null hypothesis that cell *ijk* constitutes neither a type nor an antitype. Holm's (1979) sequential test procedure was implemented, with the relaxed constraint that the first test was performed under $\alpha^* = \alpha/df$ instead of the usual $\alpha^* = \alpha/ N_c$ (see Perli et al., 1985).

The *continuity correction* turned out to be an essential component of the last simulation studies of the present authors. The correction factor, *K*, can be adjusted so that the nominal level α prevails asymptotically, that is, for large numbers of tables with the same characteristics. In contrast, Küchenhoff's (1986) continuity correction involves subtracting the constant of 0.5 from each difference between observed and expected cell frequencies. The effect of this correction is minimal for large cell frequencies, and can be dramatic for small frequencies and differences (von Eye, 2002). The continuity correction proposed by von Eye and Dunkl (1990) increases the estimate of the standard error in the denominator, e^*_{ijk}, by the factor $(e^*_{ijk} + 0.5)/(e^*_{ijk} - 0.5)$. Here again, the effect of the correction is stronger for small frequencies e^*_{ijk}. In both cases, Küchenhoff's correction and von Eye and Dunkl's correction, the magnitude of the resulting test statistic is reduced (note again, that von Eye, 2002b, pp. 71, 76, showed that Küchenhoff's correction can have the opposite effect when the difference between the observed and the expected cell frequencies is less than the correction constant).

The constant *K* is estimated iteratively. Let α be the a priori specified, nominal type I error level for the multiple level hypothesis concerning the existence of types and antitypes in a contingency table, and $\hat{\alpha}$ the estimate of the factual error level for the *M* contingency tables created in the simulation. The estimate $\hat{\alpha}$ is a function of the parameters *d*, *m*, τ_{max}, etc., but also of the table-specific constant *K*. If the simulation varies *K* while keeping all other parameters constant, the estimate becomes $\hat{\alpha} = f(K) + err$, where *f* is an unknown, typically nonlinear (and for $M \rightarrow \infty$ assumed to be monotonic and differentiable) function, and *err* is an error of unknown magnitude and distribution. The iteration attempts to estimate the equation $\hat{\alpha} = f(K)$ as precisely as possible, in spite of the error element.

The algorithm used in the present simulations employs only one α-level, α = 0.05. For this α, the constant $K = K(\alpha)$ is estimated. The estimation process itself begins by specifying the boundaries of the search interval $[K_{min}, K_{max}]$. The lower limit, K_{min}, is specified such that the estimate of α will be extremely conservative, that is, $\hat{\alpha} \rightarrow 0\%$. The upper limit, K_{max}, is specified such that the estimate of α will be extremely non-conservative, that is, $\hat{\alpha} \rightarrow 100\%$. During the iteration, this interval is reduced. Between 10 and 20 estimates $\hat{\alpha}_i(K_i)$ are identified, each of which is located close to the a priori specified nominal level α. Because of computational constraints, the number *M* of tables that are generated in the simulation is $M \leq 2000$, and because $err > 0$, an exact correspondence $\alpha = \hat{\alpha}$ is extremely unlikely. The estimator of $K(\alpha)$ is the weighted sum of the 10 to 20 K_i identified for each table. A weight of 1 is assigned if $\alpha = \hat{\alpha}_i$. Otherwise, the weights shrink exponentially with the square of the difference $\alpha - \hat{\alpha}_i$.

The five tests and test procedures used in the simulation will be described in the following paragraphs. Equations are given for the sample case of a three-dimensional table. The presentation for tables with different dimensions is straightforward. The equations given here differ from the equations given in other sources, because we indicate the location of the continuity correction, *K*.

The test of Dunkl and von Eye with Victor-expectancies $\hat{e}^*_{ijk}$ is used by the combinatoric search (CS) and by the new procedure of Lautsch and von Weber (2003)(nPr),

$$X = \frac{n_{ijk} - \hat{e}^{*}_{ijk}}{\sqrt{\sigma^{2*}_{ijk}(1-K)}},$$

with variance $\sigma^{2*}_{ijk} = \hat{e}^{*}_{ijk}(\hat{e}^{*}_{ijk} + 0.5)/(\hat{e}^{*}_{ijk} - 0.5)$. In the combinatoric search, the local test statistic X is multiplied by the antitype bonus A (see Section 2), if the average cell frequency is $m \geq 20$ and $n_{ijk} < \hat{e}_{ijk}$.

The χ-component test of Lienert (1969)(Chi),

$$\chi = \frac{n_{ijk} - \hat{e}_{ijk}}{\sqrt{e_{ijk}(1-K)}}.$$

Lehmacher's (1981) hypergeometric residual test with Küchenhoff's continuity correction, (LK),

$$z_{ijk} = \frac{n_{ijk} - \hat{e}_{ijk}}{\sqrt{N\sigma^{2}(1-K)}},$$

with

$\sigma^2 = V_{ijk} = Np_{ijk}(1 - p_{ijk} - (N - 1)(p_{ijk} - p^{*}_{ijk}))$, and $p^{*}_{ijk} = (N_{i..} - 1)(N_{.j.} - 1)(N_{..k} - 1)/(N - 1)^3$.

Küchenhoff's continuity correction subtracts 0.5 from the numerator if it is positive, and adds 0.5 to the numerator, if it is negative.

The asymptotic test of Perli, Hommel, and Lehmacher (1985)(Pe),

$$W_{ijk} = \frac{n_{ijk} - \hat{e}_{ijk}}{\sqrt{N\sigma^{2}(1-K)}},$$

with $\sigma^2 = p_{ijk}(1 + 2p_{ijk} - (p_{i..}p_{.j.} + p_{i..}p_{..k} + p_{.j.}p_{..k}))$, where this equation can be used only under simplifying assumptions.

5. Simulation results

In the following paragraphs, we present simulation results. Each set of simulations is presented in a line graph. Each connected dot represents an average of about 20.000 simulated contingency tables. Parameters under variation are number of variables (*dimension*), *d*, degrees of freedom, *df*, and average cell frequency, *m*. The graphs represent β-curves that show how the β-error decreases with *m*, for various combinations of *d* and *df*. The type II, or β-error indicates the percentage of types and antitypes placed in a table that was *not* detected. If β = 100%, none of the types or antitypes was detected (worst result). If β = 0%, all types or antitypes are detected (best result). Using the continuity correction presented by von We-

ber et al. (2003), the percentage of falsely labeled types or antitypes (α-error) was fixed to asymptotic 5%, in all runs.

To report one of the more important results right here, before the details: The new *combinatoric search method (CS)* was the only one that detected antitypes with a percentage greater than 0.1 ($\beta < 99.9$ %). It is known from empirical studies and earlier simulation results that antitypes are hard to detect when α is controlled (von Weber et al. 2003a, p. 366), but in these reports, the percentage of detected antitypes was not reported, and was small also (compare also the discussion of antitype strength in Section 3 on improvements of the F statistic). By controlling the α, we place high hurdles which can be taken only by the better performing type cells. The fact that the combinatoric search is able to detect antitypes in spite of these hurdles, is a consequence of the antitype bonus which is forcing the test statistic in the right direction.

The following graphs are presented such that the results for the tables with small numbers of variables and degrees of freedom are reported first, and *df* is the fastest varying variable.

Figure 5 shows the results for the 2-dimensional small tables with 4 or 5 degrees of freedom. For an average cell frequency of 5, the combinatoric search detects 50% of all types but no antitype. When *m* increases to 50, CS finds 73% of all types and 30% of all antitypes. Ranking the tests based on the magnitude of their β-errors, we find the rank order CS, nPr, Chi, LK, Pe. The combinatoric search is the only one able to detect antitypes under the conditions that led to the graph in Figure 5.

The results for tables with still $d = 2$ but more degrees of freedom, presented in Figures 6 and 7, emphasize the good performance of the combinatoric search. CS keeps its top position in the rank order, followed by Chi, nPr, and Pe. The last two methods perform about equally well. LK detects about as many types as CS detects antitypes. Figures 8 - 10 show the β-errors for three-dimensional tables.

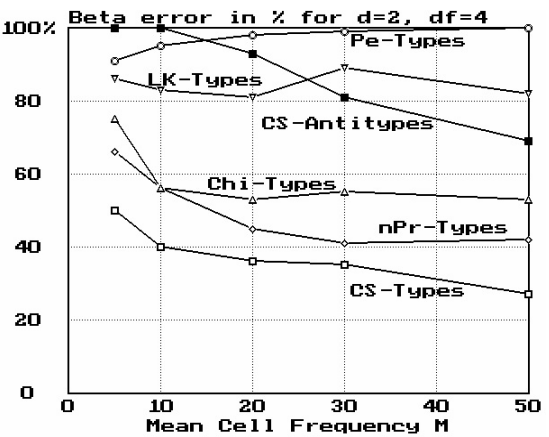

Figure 5: β-errors for $d = 2$ and $df = 4$ - 5

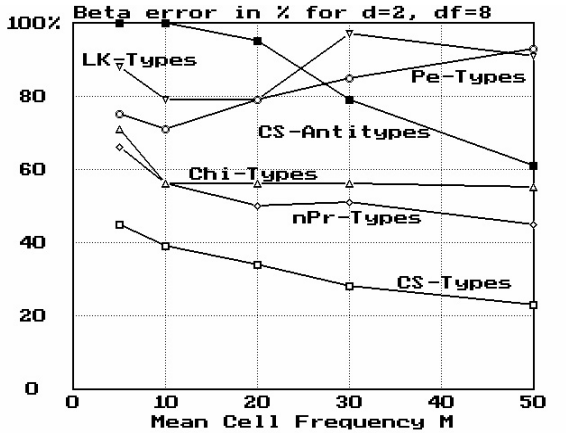

Figure 6:
β-errors for $d = 2$ and $df = 8$ - 10

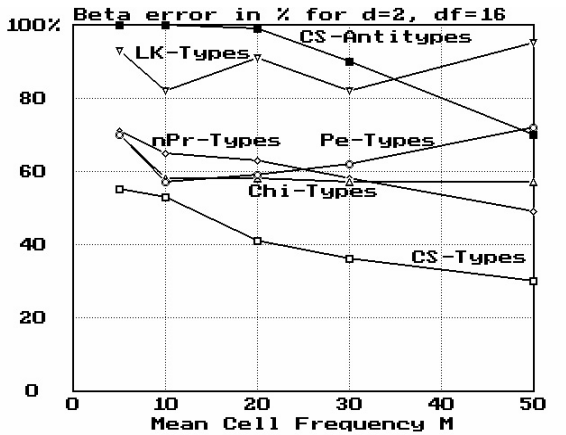

Figure 7: β-errors for $d = 2$ and $df = 16$ - 20

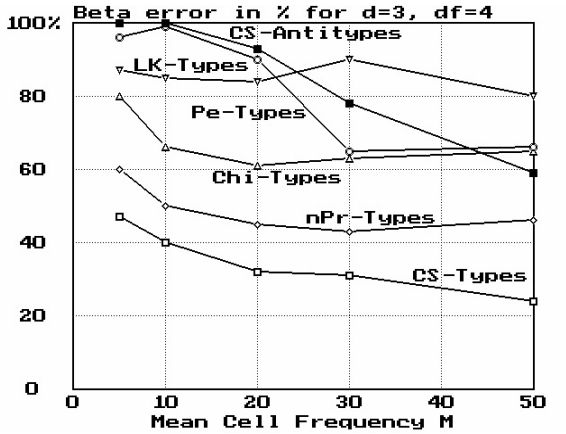

Figure 8:
β-errors for $d = 3$ and $df = 4 - 5$

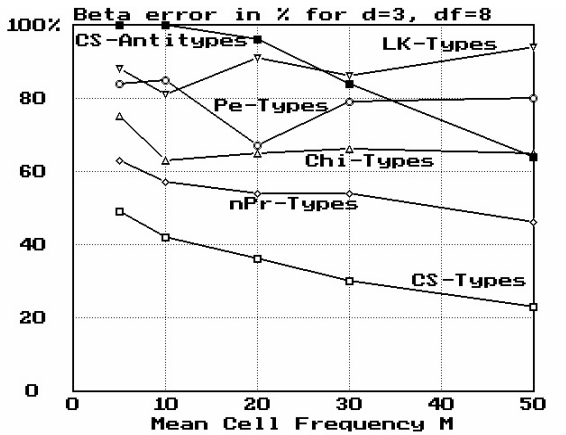

Figure 9:
β-errors for $d = 3$ and $df = 8 - 10$

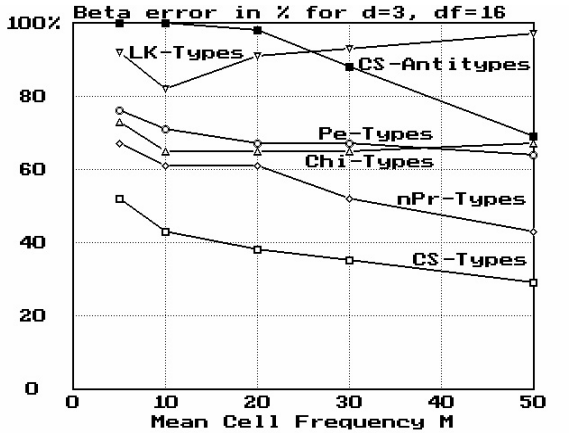

Figure 10:
β-errors for $d = 3$ and $df = 16 - 20$

Under the conditions realized for the 3-dimensional tables, the best performing test is again CS, followed by nPr, Chi, and Pe. The test at the end of the rank order is, again, LK. For larger samples, it finds even fewer types than CS finds antitypes. For $m = 50$, CS finds up to 77% of all types and up to 40% of all antitypes. Figures 11 - 13 show similar results, for four-dimensional tables.

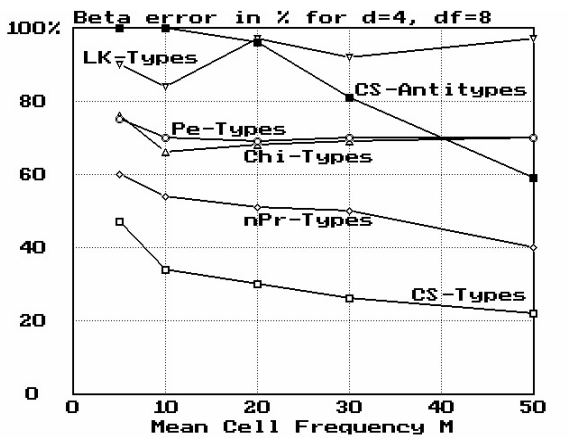

Figure 11:
β-errors for $d = 4$ and $df = 8 - 10$

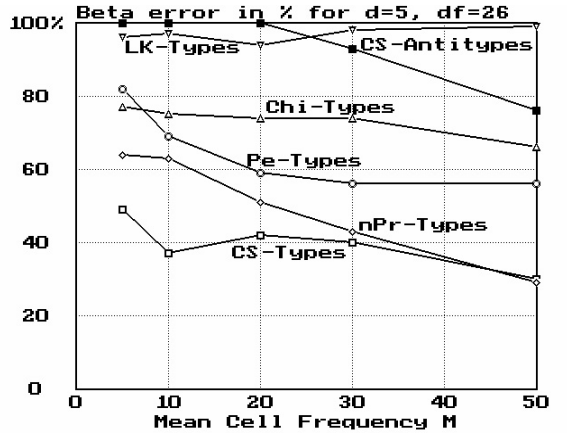

Figure 12:
β-errors for $d = 4$ and $df = 16 - 20$

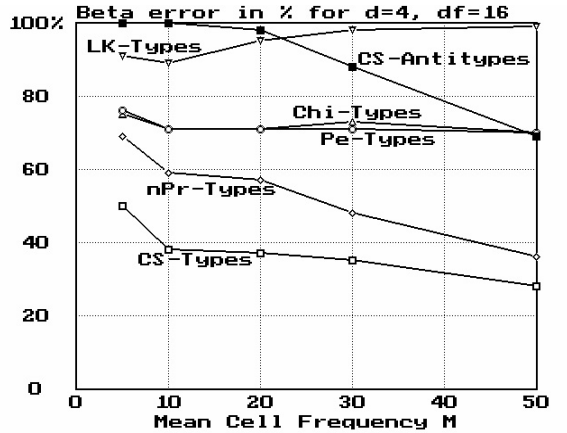

Figure 13:
β-errors for $d = 5$ and $df = 26$

A series of simulation runs was also performed for five-dimensional tables. In these runs, the rank order is CS, nPr, Pe, Chi, and LK. For $m = 50$, nPr yields the same results as CS. The probability of detecting antitypes decreases with the number of dimensions and degrees of freedom. This implies that antitypes are more likely to be detected in smaller tables with deeper dimensionality.

6. A comparison of the combinatoric search with the log-linear approach using empirical data

In this section, we re-analyze two of Kieser and Victor's (1999) examples, using the here proposed combinatoric search method. The authors (Kieser &Victor, 1999) report two examples of the use of log-linear models in the search for types and antitypes. The first involves the analysis of Lienert's (1969) LSD data. In Lienert's experiment, the incidence rates of Leuner's Psychotoxic Base Syndrome after administration of acid diethylamide (LSD) was studied in a sample of 65 participants. The reaction (Leuner, 1962) is dominated by a pattern of clouded consciousness, disturbed thinking, and altered affectivity. These 4 symptoms were scored as either present (1) or absent (2). Table 1 presents these data and results obtained with the combinatoric search, CS.

Table 1:
The combinatoric search for types and antitypes in Lienert's (1969) LSD data

```
Dimension.............          3
Degrees of freedom...          4
Maximal type strength       5.70
Alpha                       0.05
Continuity correction K  -0.4423
```

i	j	k	N_{ijk}	E_{ijk}	V_{ijk}	Tw	CEP	T/A
1	**1**	**1**	20	12.51	0.69	5.75	$<6.0\text{x}10^{-9}$	type
1	1	2	1	6.85	2.12	-0.68		
1	2	1	4	11.40	3.65	0.11		
1	2	2	12	6.24	11.24	0.15		
2	1	1	3	9.46	2.92	0.00		
2	1	2	10	5.18	8.97	0.22		
2	2	1	15	8.63	15.44	-0.08		
2	2	2	0	4.73	47.51	-4.73	$<1.3\text{x}10^{-6}$	antitype

To be able to perform the local tests, the program fixed before calculating the test statistics the Victor expectancies to values 3.0 and greater, but the originally calculated Victor expectancies are given in the Table 1. In Kieser and Victor's article, the results are essentially the same as here. However, the critical error probabilities (CEP) are smaller here. New is also, that the combinatoric search algorithm found this type-antitype pattern without postulating it as an a priori hypothesis, that is, in a fully exploratory run.

The large estimated expected cell frequency $\hat{e}^*_{222} = 47.51$ comes as a surprise, and so does the small $\hat{e}^*_{111} = 0.69$. In addition, the sum of the estimated expected cell frequencies is not equal to the observed sample size. The latter result, however, is typical of the Victor-Kieser method of analysis which does not necessarily reproduce the sample size. The reason

for this result is that the Victor-Kieser method does not need the assumption that the observed sample was drawn from the same parent population. The sample size is thus not fixed. In contrast, the sum of the expectancies from the log-linear model of variable independence must lead to $\sum_{ijk} \hat{e}^{*}_{ijk} = N$.

To illustrate, we calculate the marginal sums for the Victor-Kieser results and the classical results. We obtain $V_{1..} = 17.7$, $V_{2..} = 74.84$, $V_{.1.} = 14.7$, $V_{.2.} = 77.84$, $V_{..1} = 22.7$, and $V_{..2} = 69.84$. The total sum is $\sum_{ijk} V_{ijk} = 92.54$, a value clearly greater than the observed sample size of $N = 65$. We also calculate $\hat{V}^{*}_{111} = 17.7 * 14.7 * 22.7/92.54^2 = 0.6897$, and $\hat{V}^{*}_{222} = 74.84 * 77.84 * 69.84/92.54^2 = 47.50965$, as reported in Table 1. For the classical, first order CFA, we obtain $\sum_{ijk} \hat{e}^{*}_{ijk} = 65$. This method always reproduces the sample size because N is considered fixed.

The logical consequence of the extreme expectancies $\hat{V}_{ijk}$ is that the six remaining cells define a mean level of action. Cell (111) can thus be viewed as an outlier in the sense that the action of the drug is amplified in participants of that type to the extent that any non-affected behavior becomes impossible. Accordingly, Cell (222) is an outlier in the sense that many more participants were expected to show non-affected reactions. These two outliers show that the psychotoxic effect of LSD does not follow a simple log-linear pattern.

The second example that is re-analyzed using the CS involves data that describe a pattern of metastases in small-cell lung cancer (Kieser & Victor, 1999). Four types of chemotherapy were compared in clinical trials with a total of 1127 patients (Havemann et al., 1987; Wolf et al., 1987; Pritsch et al., 1987). It was one goal of this study to investigate whether the early formation of distant metastases in liver, skeleton, brain, and bone marrow, which is observed frequently in this type of carcinoma, occurs preferably in certain location patterns.

For the configural analyses, the four variables (1) metastases in the liver, (2) metastases in the skeleton, (3) meastases in the brain, and (4) metastases in the bone marrow were crossed. Each of the symptoms was scored as either present = 1 or absent = 2. Table 2 shows results of the combinatoric search procedure.

Kieser and Victor (1999) marked the same three cells as indicated in Table 2; however, error probabilities differed greatly. For cell (1121), their error probability is 0.022 (compared to < 0.001 as indicated by the nPr of Lautsch and von Weber, 2003). For cell (2212) Kieser and Victor indicate an error probability of 0.0016 (0.02 for the nPr), and for cell (2222) an error probability of less than 0.001 (< 0.001 for the nPr). Results suggest that the new combinatoric search detects two of the three types that had been detected by Kieser and Victor. The *New Procedure* found the same three cells, but with more emphasis on cell (1121). Based on our simulations, we trust the results created using the Combinatoric Search more than the results from the nPr.

Substantively, the results in Table 2 indicate that metastases in the brain alone are more likely than expected based on the frequency of metastases in the other three organs. The good news is that no metastases at all are far more likely than the base model would lead one to believe.

Table 2:
CS for types and antitypes in the lung cancer data

```
Dimension.............          4
Degrees of freedom...         11
Maximal type strength       3.71
Alpha                       0.05
Continuity correction K    -1.04
```

i	j	k	l	Nijkl	Eijkl	Vijkl	Tw	CEP	T/A
1	1	1	1	2	0.33	2.27	-0.24		
1	1	1	2	8	3.22	8.16	-0.03		
1	1	2	1	31	3.32	18.89	1.33	0.09	
1	1	2	2	53	32.33	67.84	-0.87		
1	2	1	1	2	1.65	3.22	-0.28		
1	2	1	2	15	16.07	11.57	0.36		
1	2	2	1	20	16.59	26.80	-0.63		
1	2	2	2	104	161.49	96.25	0.36		
2	1	1	1	3	1.25	2.12	0.00		
2	1	1	2	8	12.21	7.63	0.06		
2	1	2	1	16	12.61	17.66	-0.19		
2	1	2	2	67	122.72	63.43	0.21		
2	2	1	1	0	6.27	3.01	-0.72		
2	**2**	**1**	**2**	64	61.00	10.82	7.55	<10-13	type
2	2	2	1	31	62.98	25.05	0.46		
2	**2**	**2**	**2**	703	612.96	89.99	31.42	<10-40	type

7. A re-analysis of tables by Indurkhya and von Eye (2003), and by Lautsch (2003)

It is not surprising that two different test procedures yield different results. Sometimes, both results can be meaningful, in other cases, researchers may have to make decisions. Indurkhya and von Eye (2003) published a small frequency table with $N = 20$ and a random distribution. The results of a re-analysis using the Combinatoric Search are presented in Table 3.

The *z*-statistic used by Indurkhya and von Eye led to the identification of cell (222) as a type with the critical error probability $CEP^z = 0.0036$. Cell (221) was marked as a weak antitype with $CEP^z = 0.016$. The Combinatoric Search marks cell (221) as a strong antitype, but does not mark cell (222). What is the reason for this discrepancy? The *z* statistic was used in the context of expected cell frequencies that had been estimated under the assumption of variable independence. This implies that a surplus of cases in type cells, or a lack of cases in antitype cells is contained in the expected values. Victor's CFA model tries to identify types without this surplus and antitypes without this lack of cases. That is, Victor's model tries to find a log-linear model without residuals.

Table 3:
Re-analysis of a frequency distribution published by Indurkhya and von Eye (2003)

ijk	N_{ijk}	E_{ijk}	V_{ijk}	TW	CEP^{CS}	z	CEP^{z}
111	3	1.32	0.25	0.00		1.46	
112	0	1.08	0.05	-1.20		-1.04	
121	3	1.98	2.50	0.00		0.72	
122	0	1.62	0.46	-1.20		-1.27	
211	5	3.08	4.19	0.29		1.09	
212	0	2.52	0.77	-1.20		-1.59	
221	0	4.62	42.01	-5.24	<10-7	-2.15	0.016
222	9	3.78	7.72	0.35	0.36	2.685	0.0036

We note here again the fundamental differences between Victor's approach to CFA and the classical, Lienert approach to CFA (cf. von Eye, 2002). Victor attempts to identify that part of the observed density mass that conforms to the base model. Those parts that differ significantly can surface in the form of types and antitypes. The classic approach assumes, in its null hypothesis, that the population under study is homogeneous. Therefore, Victor's and Lienert's approaches differ in the assumptions that are made in the base model concerning the number of populations a sample was drawn from. One implication of this difference is that in the classic approach, the sample size is the same in the observed and the estimated expected frequency distributions. In Victor's approach, the estimated expected sample can differ in size from the observed sample. This was illustrated in the first empirical data example, above.

It is important to note that the number of populations considered in neither approach is clear. As soon as one type or one antitype is detected, the number of populations is greater than one. However, the reason why types or antitypes emerge can be that the cases in type or antitype cells can be mixtures from two or more populations. CFA does not ask questions concerning the number of populations. In both, Lienert's and Victor's approaches, the number of populations is assumed to be one only if there is no type and no antitype. Types and antitypes are assumed to stem from different populations than the cases in cells that conform to the base model. In Victor's CFA, expected cell frequencies are estimated under the assumption that in particular cells types or antitypes may exist. In Lienert's approach, this estimation is fully exploratory.

The last example to be given here involves a re-analysis of data that had been presented by Lautsch (2003) with a total of 1952 respondents. The author crossed the three variables (1) fear of crime, (2) risk of crime, and (3) indirect victimization. Each variable was ordinally scored at three levels. Lautsch's original analysis involved using Lehmacher's (1978) *z*-test. The search focused on types at the exclusion of antitypes. Our re-analysis is presented in the first six columns of Table 4. It is based on a multiple level of $\alpha = 0.01$ and Holm's procedure of α protection. In the last two columns, we repeat the results of the original analysis that was based on the *z*-test and the Bonferroni-adjusted $\alpha = 0.05$ of $\alpha^* = 0.025/20 = 0.00125$ for a two-sided test and the corresponding significance limit of $z = 3.05$ (Lautsch, 2000).

Table 4:
Fear of crime, risk of crime, and indirect victimization, a re-analysis of Lautsch's (2000) data

ijk	N_{ijk}	E_{ijk}	V_{ijk}	TW		Z	T/A
111	192	54.57	3.70	30.74	**type**	21.07	**type**
112	105	38.71	3.05	17.80	**type**	11.68	**type**
113	83	44.87	3.80	12.81	**type**	6.27	**type**
121	74	103.78	31.56	1.14		-3.50	
122	45	73.63	26.00	0.76		-3.84	
123	63	85.34	32.41	0.98		2.81	
131	4	68.76	21.96	-7.40	antitype	-8.98	antitype
132	5	48.78	18.09	-3.89	antitype	-6.96	antitype
133	4	56.54	22.55	-7.75	antitype	-7.86	antitype
211	20	72.03	17.21	0.20		-7.10	antitype
212	27	51.10	14.18	0.62		-3.73	
213	27	59.23	17.67	0.51		-4.72	antitype
221	209	136.99	146.87	1.84		7.70	**type**
222	154	97.19	120.99	1.08		6.90	**type**
223	167	112.65	150.82	0.47		6.23	**type**
231	40	90.77	102.17	-5.62	antitype	-6.30	antitype
232	53	64.40	84.17	-1.93		1.56	
233	62	74.64	104.92	-2.54		1.64	
311	9	58.65	14.01	-0.72		-7.34	antitype
312	4	41.61	11.54	-2.21		-6.39	antitype
313	2	48.23	14.39	-5.44	antitype	-7.40	antitype
321	72	111.54	119.59	-2.59		-4.54	antitype
322	45	79.14	98.51	-4.22	antitype	-4.46	antitype
323	63	91.72	122.81	-3.77	antitype	-3.53	
331	151	73.90	83.19	2.66		10.39	**type**
332	109	52.43	68.53	1.75		8.74	**type**
333	163	60.77	85.43	3.00		14.91	**type**

Lautsch's data are unusual because of the large number of types and antitypes found by both approaches. It seems that there is no simple base model for these data. Another unusual aspect is that the number of types and antitypes found by the two approaches differs greatly. Each of the types and antitypes found by the combinatoric search was also found by the classic method. However, the classic approach marks all cells as types or antitypes, with the exception of three. If a stricter significance level is used, the number of unmarked cells increases only to seven.

8. Discussion and recommendations

This article presents the first application of the Combinatoric Search Procedure for CFA that had been proposed by Kieser and Victor (1991). This procedure works convincingly well. The direct comparison with four other tests showed that the Combinatoric Search Procedure is better than the comparison methods in all cases that were studied. In the cases of small mean frequencies ($m = 5$), the Combinatoric Search detects about 50% of all types, whereas the other methods find only 30-40%. In the cases of mean cell frequencies of $m = 20$ or $m = 30$, the Combinatoric Search is performing about 10 to 20% better than the next method. An exception we find for tables with high dimensionality ($d = 5$). Here, the New Procedure (nPr) is performing at the same level. For well-occupied tables ($m = 50$) the Combinatoric search detects about 70% of all types, this is about 20% more than the other methods. Here again, the New Procedure performs at the same level as the Combinatoric Search, but only in cases of high dimensionality. The Combinatoric Search finds, in addition, about 30-40% of all antitypes in tables with $m = 50$. Only in the case of high dimensionality ($d = 5$), this value is reduced to 25% detected antitypes.

Most of the improvements over the other procedures are due to the inclusion of additional terms in the search statistic. These terms are new. They lead to an improvement in the detection of types and antitypes in tables with relatively large observed cell frequencies and, in particular, in tables with three or more dimensions.

Another result of the present research is the ranking of the five tests included in the simulations. The Combinatoric Search Procedure was number one under all conditions, followed by Lautsch and von Weber's (2000) New Procedure. A tie was found between Lienert's X^2-test and Perli et al.'s w-test. Whereas the w-test seems to perform better if the table is spanned by more than two variables, in lower-dimensional tables and in tables with fewer degrees of freedom, the X^2-test works better.

The performance of the Combinatoric Search Procedure comes with a price. It is very computer-intensive. 99.99% of the computational time that is used by this procedure is needed for the estimation of the continuity correction constant K. This constant is important because it helps keep the error probability α at the a priori determined level. Without K, we have no indication as to whether a test suggests conservative or non-conservative decisions, and there is no kit to repair this problem. Future research will focus on devising more computer-efficient methods of estimating K. The current algorithm needs about 20.000 simulated tables when $\alpha = 0.05$. We hope to reduce this number without loss of accuracy. In addition, we hope that the increasing speed of computers will help minimize this problem.

References

1. Dunkl, E., von Eye, A. (1990). Kleingruppentests gegen Victor-Typen und -Syndrome. Zeitschrift für Klinische Psychologie, Psychopathologie und Psychotherapie, 44, 46-51.
2. Goodman, L. A. (1968). The analysis of cross-classified data: Independence, quasi-independence and interactions in contingency tables with or without missing entries. Journal of the American Statistical Association, 324, 1091-1131.
3. Goodman, L. A. (1991). Measures, models, and graphical displays in the analysis of cross-classified data. Journal of the American Statistical Association, 86, 1085-1111.

4. Holm, S. (1979). A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics, 6, 65-70.
5. Kieser, M., Victor, N. (1991). A test procedure for an alternative approach to configural frequency analysis. Methodika, 5, 87-97.
6. Kieser, M., Victor, N. (1999). Configural Frequency Analysis (CFA) Revisited – a New Look at an Old Approach. Biometrical Journal, 41, 967-983.
7. Küchenhoff, H. (1986). A note on a continuity correction for testing in three-dimensional configural frequency analysis. Biometrical Journal, 28, 465-468.
8. Lautsch, E., von Weber, S. (2003). Eine neue Testprozedur in der KFA. Psychology Science, 45, 389-399.
9. Lehmacher (1981). A more powerful simultaneous test procedure in configural frequency analysis. Biometrical Journal, 23, 429-436.
10. Lienert, G.A. (1969). Die Konfigurationsfrequenzanalyse als Klassifikationsmittel in der klinischen Psychologie. In: Irle, M. (Hrsg.), Bericht über den 26. Kongreß der Deutschen Gesellschaft für Psychologie 1968 in Tübingen (pp. 244 - 255). Göttingen: Hogrefe.
11. Perli (1985). Testverfahren in der Konfigurationsfrequenzanalyse. Erlangen: Palm und Enke.
12. 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.
13. Victor, N. (1983). A note on contingency tables with one structural zero. Biometrical Journal, 25, 283-289.
14. Victor, N. (1989). An alternative approach to configural frequency analysis. Methodika, 3, 61-73.
15. von Eye, A., Spiel, C., & Rovine, M. J. (1995). Concepts of nonindependence in Configural Frequency Analysis. Journal of Mathematical Sociology, 20, 41-54.
16. von Eye, A. (2002). Configural Frequency Analysis - Methods, Models, and Applications. Mahwah, NJ: Lawrence Erlbaum.
17. von Weber, S. (2000). Ein Vergleich in der KFA verwendeter Tests mittels Simulationsrechnungen. Psychologische Beiträge, 42, 260-272.
18. von Weber, S., Lautsch, E., von Eye, A. (2003a). Table-specific continuity corrections for Configural Frequency Analysis. Psychology Science, 45, 355-368.
19. von Weber, S., Lautsch, E., von Eye, A. (2003b). On the limits of CFA: Analyzing small tables. Psychology Science, 45, 339-354.
20. von Weber, S., von Eye, A, Lautsch, E. (2004). Error type II measures for the analysis of 2 x 2 tables. Understanding Statistics (in press).