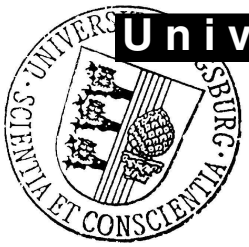

Universität Augsburg

Institut für
Mathematik

Ali Ünlü, Anatol Sargin

Maximum likelihood methodology for *diff* fit measures for quasi orders

Preprint Nr. 30/2008 — 12. November 2008
Institut für Mathematik, Universitätsstraße, D-86135 Augsburg http://www.math.uni-augsburg.de/

Impressum:

Herausgeber:
Institut für Mathematik
Universität Augsburg
86135 Augsburg
`http://www.math.uni-augsburg.de/forschung/preprint/`

ViSdP:
Ali Ünlü
Institut für Mathematik
Universität Augsburg
86135 Augsburg

Maximum likelihood methodology for *diff* fit measures for quasi orders

Ali Ünlü *, Anatol Sargin

University of Augsburg, D-86135 Augsburg, Germany

Abstract

Three inductive item tree analysis algorithms have been proposed for deriving quasi orders from dichotomous data. These procedures have been treated descriptively, without examining theory. In this paper, we introduce maximum likelihood methodology for the inductive item tree analysis methods. The *diff* fit measures of these methods can be interpreted as maximum likelihood estimators. We show that the estimators are asymptotically efficient, and hence they are asymptotically normal, asymptotically unbiased, and consistent. In simulation studies, the algorithms are compared regarding finite sample consistency, population ranks, and population symmetric differences. The approach to fit measures presented in this paper can be applied to any, sufficiently smooth, coefficient for multinomial count data. In particular, it allows introducing maximum likelihood methodology for measures assessing the fit of general knowledge structures.

Key words: Inductive item tree analysis, *diff* fit measures, knowledge space theory, quasi order, multinomial count data, maximum likelihood, asymptotics

1 Introduction

Deriving quasi orders (reflexive and transitive binary relations) on sets of dichotomous items plays an important role in knowledge space theory (KST) (Albert & Lukas (1999); Doignon & Falmagne (1999)), introduced in Section 2.

* Corresponding author. Phone: +49 821 598 2236; fax: +49 821 598 2200.
Email addresses: ali.uenlue@math.uni-augsburg.de (Ali Ünlü), anatol.sargin@math.uni-augsburg.de/ (Anatol Sargin).
URLs: http://www.math.uni-augsburg.de/∼uenlueal/ (Ali Ünlü), http://stats.math.uni-augsburg.de/mitarbeiter/sargin/ (Anatol Sargin).

In Sargin & Ünlü (2008), we correct and improve the original version of inductive item tree analysis (IITA) (Schrepp (1999, 2003)), which is a data-analytic method for building quasi orders. That paper introduces two new algorithms, corrected and minimized corrected IITA, and compares these three algorithms in a simulation study. It is seen that, on average, the corrected and minimized corrected versions perform better than the original, in terms of both smaller *dist* values and numbers of erroneously detected implications.
So far, these IITA procedures have been treated descriptively, without examining theory. In the present paper, we focus on theoretical aspects and further comparisons of the three algorithms. We demonstrate that the *diff* coefficients of these methods can be transformed into maximum likelihood estimators, by division through the square of sample size. We discuss the quality properties asymptotic efficiency, asymptotic normality, asymptotic unbiasedness, and consistency. Consistency in finite samples is illustrated in simulation, and comments on the use of asymptotic normality in practice are given. Remarkably, the approach of this paper is a general one, in the sense that it can be applied to any, sufficiently smooth, coefficient for multinomial count data. In Section 6, we briefly exemplify that with the discrepancy index, which can be used as a measure for general knowledge structures.
Theoretical considerations in population quantities are important. Supposing the population completely to be known seems to be the way to begin in constructing sound fit measures for quasi orders. After providing justification for a measure in a known population, one should consider sampling problems concerning estimation and testing (Goodman & Kruskal (1979)). Literature on the IITA algorithms, however, has dealt with samples rather than the population.
In this paper, we discuss population analogs of the *diff* coefficients, and we propose evaluating these fit measures based on rank ordered population values. This approach is justified given the consistency property of the estimators. We compare the IITA algorithms in extensive simulation studies regarding finite sample consistency, population ranks for the *diff* fit measures, and population symmetric differences of the true and obtained quasi orders and their corresponding knowledge structures. In almost all cases, the original IITA version is clearly outperformed by the two new versions; this confirms the results obtained in Sargin & Ünlü (2008). We close with a discussion containing a summary, concluding remarks, and some suggestions for future research.

2 Basic concepts of knowledge space theory

This section reviews some of the basic deterministic and probabilistic concepts of KST (Albert & Lukas (1999); Doignon & Falmagne (1999)), which are relevant for this work.

2.1 Deterministic concepts

Assume a set Q of n dichotomous items. Mastering an item $j \in Q$ may imply mastering another item $i \in Q$. If no response errors are made, these implications, $j \rightarrow i$, entail that only certain response patterns are possible. Those response patterns are called knowledge states, and the set of all knowledge states (including $\emptyset$ and Q), $\mathcal{K}$, is called a knowledge structure. Note that $\mathcal{K}$ is a subset of 2^Q, where 2^Q is the power set of Q. When a knowledge structure is closed under union, it is called a knowledge space. Knowledge spaces closed under intersection are called quasi ordinal. Implications are assumed to form a quasi order on the item set Q. Quasi orders are referred to as surmise relations in KST, and they bijectively correspond to quasi ordinal knowledge spaces (Doignon & Falmagne (1999)).
An application of these concepts is, for example, an aptitude test, where people can solve or fail to solve a question. Throughout, we use this interpretation to illustrate the algorithms.

2.2 Probabilistic concepts

Implications are latent and not directly observable, due to random response errors. A person who is actually unable to solve an item, but does so, makes a lucky guess. On the other hand, a person makes a careless error, if he fails to solve an item which he is capable of mastering. A probabilistic extension of the knowledge structure model covering random response errors is the basic local independence model in KST.

A quadruple $(Q, \mathcal{K}, p, r)$ is called a basic local independence model (BLIM) (Doignon & Falmagne (1999)) if and only if

(1) $(Q, \mathcal{K})$ is a knowledge structure,
(2) p is a probability distribution on $\mathcal{K}$, i.e., $p : \mathcal{K} \rightarrow]0,1[\,, K \mapsto p(K)$, with $p(K) > 0$ for any $K \in \mathcal{K}$, and $\sum_{K \in \mathcal{K}} p(K) = 1$,
(3) r is a response function for $(Q, \mathcal{K}, p)$, i.e., $r\colon\ 2^Q \times \mathcal{K} \rightarrow [0,1]$, $(R, K) \mapsto r(R, K)$, with $r(R, K) \geq 0$ for any $R \in 2^Q$ and $K \in \mathcal{K}$, and $\sum_{R \in 2^Q} r(R, K) = 1$ for any $K \in \mathcal{K}$,
(4) r satisfies local independence, i.e.,

$$r(R,K) = [\prod_{q \in K \setminus R} \beta_q] \cdot [\prod_{q \in K \cap R} (1-\beta_q)] \cdot [\prod_{q \in R \setminus K} \eta_q] \cdot [\prod_{q \in Q \setminus (R \cup K)} (1-\eta_q)],$$

with two constants $\beta_q, \eta_q \in [0,1[$ for each $q \in Q$, respectively called careless error and lucky guess probabilities at q.

To each state $K \in \mathcal{K}$ is attached a probability $p(K)$ measuring the likelihood that an examinee is in state K (point 2). For $R \in 2^Q$ and $K \in \mathcal{K}$, $r(R, K)$ specifies the conditional probability of response pattern R for an examinee in state K (point 3). The item responses of an examinee are assumed to be independent given the knowledge state of the examinee. The response error probabilities β_q, η_q $(q \in Q)$ are attached to the items and do not vary with the knowledge states (point 4).
The resulting probability distribution on the set of all response patterns is

$$\rho(R) = \sum_{K \in \mathcal{K}} r(R, K) p(K).$$

We consider a random sample of size m. The data are the absolute counts $m(R)$ of response patterns $R \in 2^Q$, i.e., $\mathbf{x} = (m(R))_{R \in 2^Q}$. The examinees are assumed to give their responses independent of each other. The true probability of occurence $\rho(R)$ of any response pattern R is assumed to stay constant across the examinees, and to be strictly larger than zero. Then the data $\mathbf{x}$ are the realization of a random vector, $\mathbf{X} = (X_R)_{R \in 2^Q}$, which is distributed multinomially over 2^Q. In other words, the probability of observing the data $\mathbf{x}$, i.e., the realizations $X_R = m(R)$, is

$$\begin{aligned} \mathbb{P}(\mathbf{X} = \mathbf{x}) &= \mathbb{P}(X_\emptyset = m(\emptyset), \ldots, X_Q = m(Q)) \\ &= \frac{m!}{\prod_{R \in 2^Q} m(R)!} \prod_{R \in 2^Q} \rho(R)^{m(R)}, \end{aligned}$$

where $\rho(R) > 0$ for any $R \in 2^Q$, $\sum_{R \in 2^Q} \rho(R) = 1$, and $0 \leq m(R) \leq m$ for any $R \in 2^Q$, $\sum_{R \in 2^Q} m(R) = m$.

3 IITA algorithms

The three IITA algorithms are exploratory methods for extracting surmise relations from data. In each algorithm, competing binary relations are generated (in the same way for all three versions), and a fit measure is computed for every relation in order to find the quasi order that fits the data best. In the following, the algorithms are briefly reviewed and their differences pointed out.

3.1 Generating the selection set

The first step for all three algorithms is the inductive generation of surmise relations. For two items i and j, $b_{ij} := |\{R \in D | i \notin R \wedge j \in R\}|$ is the number

of counterexamples, that is, the number of observed response patterns in the data matrix D contradicting $j \to i$. Based on these values, binary relations $\sqsubseteq_L$ for $L = 0, ..., m$ are defined. Let $i \sqsubseteq_0 j :\Leftrightarrow b_{ij} = 0$. The relation $\sqsubseteq_0$ is transitive, and based on that, all the other transitive relations $\sqsubseteq_L$ are constructed inductively.
Assume $\sqsubseteq_L$ is a transitive relation. Define the set $S_{L+1} := \{(i,j)|b_{ij} \leq L + 1 \wedge i \not\sqsubseteq_L j\}$. This set consists of all item pairs that are not already contained in the relation $\sqsubseteq_L$ and have at most $L+1$ counterexamples. From these item pairs those are excluded that cause an intransitivity in $\sqsubseteq_L \cup S_{L+1}$, and the remaining item pairs are referred to as $S^{(1)}_{L+1}$. This process continues iteratively, say k times, until no intransitivity is caused anymore. The generated relation $\sqsubseteq_{L+1} := \sqsubseteq_L \cup S^{(k)}_{L+1}$ is then transitive by construction. Because $\sqsubseteq_0$ is reflexive, all generated relations are. Hence $\sqsubseteq_L$ is a quasi order for every $L = 0, ..., m$.

Besides the construction of the quasi orders, it is very important to find that quasi order which fits the data best. In all algorithms, the idea is to estimate the numbers of counterexamples for each quasi order, and to find, over all competing quasi orders, the minimum value for the discrepancy between the observed and expected numbers of counterexamples.

3.2 Determining the best fitting quasi order

Let $p_i := |\{R \in D|i \in R\}|/m$ be the relative solution frequency of an item i. A violation of an underlying implication is only possible due to random errors. To compute the expected numbers of counterexamples, b^*_{ij}, error probabilities are needed. In all three algorithms, the error probabilities are assumed to be equal for all items. In the original and corrected IITA versions, that single error rate is estimated by

$$\gamma_L := \frac{\sum\{b_{ij}/(p_j m)|i \sqsubseteq_L j \wedge i \neq j\}}{(|\sqsubseteq_L| - n)},$$

where $|\sqsubseteq_L| - n$ is the number of non-reflexive item pairs in $\sqsubseteq_L$.
In the minimized corrected IITA version, minimizing the *diff* coefficient as a function of the error probability γ_L gives $\gamma_L = -\frac{x_1+x_2}{x_3+x_4}$, where

$$x_1 := \sum_{i \not\sqsubseteq_L j \wedge j \sqsubseteq_L i} -2b_{ij}p_i m + 2p_i p_j m^2 - 2p_i^2 m^2,$$
$$x_2 := \sum_{i \sqsubseteq_L j} -2b_{ij}p_j m,$$

$$x_3 := \sum_{i \not\sqsubseteq_L j \wedge j \sqsubseteq_L i} 2p_i^2 m^2,$$
$$x_4 := \sum_{i \sqsubseteq_L j} 2p_j^2 m^2.$$

For details on the derivation, see Sargin & Ünlü (2008).
Under every relation, the algorithms compute the expected numbers of counterexamples for each (non-reflexive) item pair. First, we present the estimators used in the original IITA algorithm. If the relation $\sqsubseteq_L$ provides an implication $j \to i$, the expected number of counterexamples is computed by $b_{ij}^* = p_j \gamma_L m$. If $(i,j) \notin \sqsubseteq_L$, no dependency between the two items is assumed, and $b_{ij}^* = (1-p_i)p_j m(1-\gamma_L)$. In this formula, $(1-p_i)p_j m$ is the usual probability for two independent items, and the factor $1-\gamma_L$ is assumed to state that no random error occurred. For criticism on these estimates, see Sargin & Ünlü (2008).
For the corrected and minimized corrected IITA versions, the same estimates are used.

- If $(i,j) \in \sqsubseteq_L$, set $b_{ij}^* = p_j \gamma_L m$.
- If $(i,j) \notin \sqsubseteq_L$ and $(j,i) \notin \sqsubseteq_L$, set $b_{ij}^* = (1-p_i)p_j m$.
- If $(i,j) \notin \sqsubseteq_L$ and $(j,i) \in \sqsubseteq_L$, set $b_{ij}^* = (p_j - p_i + p_i \gamma_L)m$.

Motivation for and derivation of these estimates can be found in Sargin & Ünlü (2008).

Three measures for the fit of each relation $\sqsubseteq_L$ to the data matrix D are the *diff* coefficients defined by

$$\mathit{diff}(\sqsubseteq_L, D) := \sum_{i \neq j} \frac{(b_{ij} - b_{ij}^*)^2}{n(n-1)},$$

where corresponding estimates b_{ij}^* are used. They give the average sums of the quadratic differences between the observed and expected numbers of counterexamples under the relation $\sqsubseteq_L$. The smaller the *diff* values are, the better is the fit of the relation to the data. Therefore, the IITA algorithms look for the smallest values of the *diff* coefficients and return the corresponding quasi orders.

4 Maximum likelihood methodology

In this section, we introduce the population analogs of the *diff* fit measures, interpret the coefficients as maximum likelihood estimators for the corresponding population values, and show for these estimators the quality properties

asymptotic efficiency, asymptotic normality, asymptotic unbiasedness, and consistency. The use of asymptotic normality in practice is further commented on in Section 6.

4.1 The diff *coefficients as maximum likelihood estimators*

Consider the transformed sample *diff* coefficients $\mathit{diff}_t := \mathit{diff}/m^2$. The division is necessary to cancel out sample size m in replacements of sample quantities with population quantities. Given the multinomial probability distribution on the set of all response patterns, make the following replacements in the arguments, b_{ij} and p_i, of the sample diff_t coefficients:

$$\frac{b_{ij}}{m} \to \mathbb{P}(i=0, j=1) = \sum_{R \in 2^Q, i \notin R \wedge j \in R} \rho(R),$$
$$p_i \to \mathbb{P}(i=1) = \sum_{R \in 2^Q, i \in R} \rho(R).$$

This gives three population diff_t coefficients corresponding to the sample diff_t coefficients.
The sample diff_t coefficients are the obvious sample analogs of these population fit measures. They are reobtained by replacing the arguments $\rho(R)$ of the population diff_t measures with the maximum likelihood estimates $m(R)/m$ of the multinomial distribution. According to the invariance property of maximum likelihood estimation, the sample diff_t coefficients are the maximum likelihood estimators for the corresponding population diff_t coefficients. The invariance property states that if $\hat{\theta}$ is the maximum likelihood estimator for θ, then for any function $f(\theta)$, the maximum likelihood estimator for $f(\theta)$ is $f(\hat{\theta})$ (Casella & Berger (2002); Zehna (1966)).

4.2 Asymptotic properties of the diff_t *coefficients*

Next we present an application of established maximum likelihood asymptotics. Though this is a straightforward application, it is novel and important in the so far ad-hoc discussion of data-analytic methods in KST. Since the following techniques are well-known, the explanations are kept succinct. For technical details on asymptotic properties and regularity conditions, see Bishop, Fienberg, & Holland (1975), Casella & Berger (2002), and Witting & Müller-Funk (1995).
Maximum likelihood estimators possess a number of asymptotic quality properties, given certain regularity conditions are satisfied. Important properties

are asymptotic efficiency (the most precise estimates are produced), and implied by this property, asymptotic normality, asymptotic unbiasedness (estimates converge in expectation to the true values), and consistency (estimates converge in probability to the true values). It can be verified that the maximum likelihood estimator for the multinomial distribution fulfills required regularity conditions and hence is asymptotically efficient.
The population $diff_t$ coefficients are continuous functions of the multinomial cell probabilities $\rho(R)$. [1] Therefore the corresponding sample $diff_t$ coefficients are asymptotically efficient, asymptotically normal, asymptotically unbiased, and consistent estimators for the population values.

4.3 Illustrating consistency

One possibility to assess and compare the quality of asymptotic properties for finite samples for the three IITA algorithms is by simulation. We exemplify that with the consistency property. First, we visually illustrate consistency using one quasi order. Theoretically, consistency is formulated and holds for any single quasi order. The rate of convergence may vary from quasi order to quasi order. Second, to get a rough structure-independent evaluation, we aggregate the results obtained for 100 quasi orders.
In this paper, the performances of the original, corrected, and minimized corrected IITA algorithms are compared in simulation studies. Throughout, simulations were realized using the R statistical computing environment (R Development Core Team (2006); http://www.r-project.org/). The source files are freely available from the authors.
The simulation study illustrating consistency is based on nine items and is as follows.

(1) All reflexive pairs are always added to the relation $\mathcal{R}$. A constant δ is set randomly (Sargin & Ünlü (2008)), which gives the probability for adding each of the remaining 72 item pairs to the relation. The transitive closure $\sqsubseteq$ of this relation $\mathcal{R}$ is computed, and is the underlying (true) quasi order.
(2) Fifty data matrices are simulated for each of the increasing sample sizes 100, 1000, 10000, and 25000 in the following way. From the set $\{K \in 2^Q : (i \sqsubseteq j \wedge j \in K) \rightarrow i \in K\}$ of all response patterns consistent with $\sqsubseteq$ an element is drawn randomly. For this drawn pattern all entries are changed from 1 to 0 or from 0 to 1, with a same prespecified error probability τ. This is simulating with a special case of the BLIM.
(3) Under all three algorithms, the sample and population $diff_t$ coefficients are computed.

[1] Note that $\rho(R) > 0$ for all response patterns $R \in 2^Q$. This assumption is essential for assuring continuity of the population $diff_t$ coefficients.

Figure 1 gives a graphical display of consistency for one quasi order (for $\tau = 0.10$); running the previous three simulation steps once.

[Insert Figure 1 about here]

Figure 1 shows boxplots for all three IITA algorithms, within each of the sample sizes of the 50 computed sample $diff_t$ values. The three population $diff_t$ values are shown as horizontal lines in the plots. This graphic illustrates that the population values are better attained and the sample values are less dispersed with increasing sample size, for all three algorithms. The results are better for the corrected and minimized corrected IITA versions than for the original. The corrected and minimized corrected algorithms have a higher speed of convergence. In particular, they achieve the population values with a much higher accuracy than the original algorithm, which shows, even for a sample size of 25000, clear discrepancies between sample and population values. Hence consistency, which is guaranteed by theory, manifests in smaller finite sample sizes for the two new IITA versions.

Table 1 summarizes the aggregated results obtained for 100 quasi orders (for $\tau = 0.10$); running the three simulation steps 100 times.

[Insert Table 1 about here]

Table 1 shows, for each combination of ϵ (0.01, 0.001, 0.0001) and sample size, the relative frequencies of 5000 data matrices satisfying $|\hat{\theta}_n - \theta| > \epsilon$, where $\hat{\theta}_n$ and θ stand for the sample and population $diff_t$ coefficients, respectively. The entries represent estimates of the probabilities $\mathbb{P}(|\hat{\theta}_n - \theta| > \epsilon)$ used in the definition of consistency, where the probability is taken with respect to the true multinomial distribution. For instance, the first entry says that the probability for obtaining a sample $diff_t$ value, for a sample size of 100, differing more than 0.01 from the population $diff_t$ value is, approximately, 0.0010. This is on average, independent of the underlying quasi order.
Under all three algorithms, for each ϵ, the relative frequencies are decreasing with increasing sample size (except for one case, mentioned below). Again, the two new IITA versions outperform the original. The original version shows the lowest speed of convergence, and for $\epsilon = 0.0001$, from sample sizes 10000 to 25000, the relative frequency is even increasing. The corrected and minimized corrected IITA algorithms perform well and quite similar, with a slight advantage for the minimized corrected.

In sum, we have seen that the *diff* coefficients of the IITA algorithms can be interpreted as maximum likelihood estimators possessing desirable asymptotic properties. Based on the consistency property, next we propose evaluating the *diff* fit measures via rank ordered population values.

5 Comparisons of the three algorithms

In prior publications, only sample, not population, quantities have been considered. The simulation study in this section is theoretical, in the sense of solely dealing with values for a known population. The following summary statistics (evaluation criteria) are investigated in population, not sample, quantities.
The symmetric difference, at the level of items (*dist*), of the obtained and underlying quasi orders is used as a distance measure. Since there is a bijection between quasi orders and their corresponding knowledge structures, the symmetric difference can also be considered at the level of knowledge states ($dist^\star$). The results obtained at the two levels may differ; for example, the original IITA algorithm may have moderately lowest *dist* but considerably highest $dist^\star$ values (see Table 2). Therefore we introduce the rank statistic (*rk*) as a third useful measure. Given a set of competing quasi orders, which is required to include the underlying one, this statistic computes the rank of the true quasi order in the ordered list of population $diff_t$ values.
This population based approach is justified according to the asymptotic theory discussed in Section 4. The sample $diff_t$ values converge in probability (and expectation) to the population $diff_t$ values.

5.1 Procedure of the simulation study

In the simulation study nine items are used. The general simulation scheme consists of five parts. First, the underlying quasi order is generated randomly. Second, the set of competing quasi orders is constructed according to the inductive procedure of the IITA algorithms. Third, the underlying quasi order is added to the selection set. Fourth, the population $diff_t$ coefficients are computed. Fifth, the three algorithms are compared regarding symmetric differences and ranks. More precisely:

(1) All reflexive pairs are always added to the relation $\mathcal{R}$. A constant δ is set randomly (Sargin & Ünlü (2008)), which gives the probability for adding each of the remaining 72 item pairs to the relation. The transitive closure $\sqsubseteq$ of this relation $\mathcal{R}$ is computed, and is the underlying quasi order.
(2) To generate a selection set of quasi orders, a binary 5000×9 data matrix is simulated. From the set $\{K \in 2^Q : (i \sqsubseteq j \wedge j \in K) \rightarrow i \in K\}$ of all response patterns consistent with $\sqsubseteq$ an element is drawn randomly. For this drawn pattern all entries are changed from 1 to 0 or from 0 to 1, with a same prespecified error probability τ. The inductive construction procedure is applied to the simulated data matrix. [2]

[2] The idea is to obtain a large as possible number of quasi orders in the selection set. Experimentation (not reported here) has shown that for sample sizes greater

(3) If the underlying quasi order $\sqsubseteq$ is not contained in the selection set, it is added.
(4) Under all three algorithms, the population $diff_t$ coefficients are computed for all quasi orders of the selection set.
(5) The three algorithms are compared with respect to three criteria: the symmetric differences *dist* and $dist^\star$ of the obtained (with smallest population $diff_t$ value) and underlying quasi orders and corresponding knowledge structures, respectively, and the rank *rk* of the underlying quasi order among the population $diff_t$ values.

The error probabilities take the values 0.03, 0.05, 0.08, 0.10, 0.15, and 0.20. For each of these error probabilities, the previous five simulation steps are run 1000 times.

5.2 Results of the simulation study

For each of the three algorithms, for every error probability, three population summary statistics are computed. They are the means of the 1000 *dist*, $dist^\star$, and *rk* values. These summary statistics are reported in Table 2.

[Insert Table 2 about here]

Table 2 shows the following results:

(1) Summary statistic *dist*: For the small error rates 0.03 and 0.05, the original algorithm gives better average *dist* results than the corrected and minimized corrected. For all other τ values, the two new versions perform clearly better than the original. This is especially the case for the large error probabilities 0.15 and 0.20.
The average population *dist* values show a similar pattern as the average sample *dist* values reported in Sargin & Ünlü (2008). Those descriptive results hence are substantiated through theoretical considerations. In both simulation studies, the two new versions outperform the original, yet the difference in performance is larger in terms of population quantities.
For any τ value, the minimized corrected IITA algorithm performs slightly better than the corrected. This shows that the minimized corrected version is better asymptotically.
(2) Summary statistic $dist^\star$: For all error probabilities, the average $dist^\star$ statistic gives the same ranking; listed from worst to best, original, corrected, and minimized corrected IITA. The results are quite similar for the corrected and minimized corrected algorithms. Compared to the original

than 5000 barely any improvement of the selection set is achieved. Sample sizes smaller than 5000 have led to smaller selection sets.

version, the corrected and minimized corrected IITA algorithms perform very well. For error probabilities up to 0.10, their average $dist^\star$ values are smaller than 3. The original version, however, shows a bad performance already for $\tau = 0.05$. The results strongly worsen, reaching a maximum average $dist^\star$ value of 174.80 for $\tau = 0.20$. For the corrected and minimized corrected versions, the corresponding average $dist^\star$ values are 14.76 and 10.81, respectively.

(3) Summary statistic *rk*: For all error probabilities, the average *rk* statistic gives the same ranking; listed from worst to best, original, corrected, and minimized corrected IITA. The corrected and minimized corrected IITA algorithms perform quite similar. Compared to the original version, they produce good results, especially for larger error rates. For $\tau = 0.20$, the corrected and minimized corrected versions give average *rk* values of 4.79 and 3.86, respectively, while the original algorithm shows a considerably larger average *rk* value of 16.96.

Some remarks are in order regarding the results of the simulation study.

(1) Overall, the minimized corrected version performs best, second comes the corrected, and worst is the original (with respect to all three summary statistics). We have obtained similar results for the two new algorithms. For each of the three summary statistics, the original version has shown considerably bad results for larger error probabilities.

(2) Further analyses made using ranks (of underlying quasi orders) show that the original version, compared to the other two algorithms, not only performs worse based on average ranks, but also has higher maximum ranks. For every error probability, the maximum of the 1000 *rk* values is greater. For instance, we obtained the maximum ranks 22, 7, 6 (for $\tau = 0.03$) and 40, 31, 15 (for $\tau = 0.10$) for the original, corrected, and minimized corrected algorithms. Moreover, the original version is outperformed concerning the number of *rk* values that are at most as large as 3 (first three ranks). For instance, we obtained the first three ranks 893, 940, 960 times (for $\tau = 0.03$) and 645, 830, 919 times (for $\tau = 0.10$) for the original, corrected, and minimized corrected algorithms. These summary statistics measure rank variability and show that the original IITA algorithm has a wider range for the *rk* values.

(3) That the original algorithm gives better average *dist* results in population quantities for the error probabilities 0.03 and 0.05 can be explained in the same way as we did for sample quantities in Sargin & Ünlü (2008). The incorrect estimation scheme of the original algorithm produces good results specifically when the size of the underlying quasi order is large. For a large quasi order $\sqsubseteq$, there are predominantly the cases $i \sqsubseteq j$, for which correct estimators are used. For the cases $i \not\sqsubseteq j$, however, incorrect estimators are applied, and the discrepancies between the observed and expected numbers of counterexamples are large. This implies that, for

an underlying large quasi order, the $diff_t$ values for small quasi orders of the selection set are large (pulling apart the $diff_t$ value for the true quasi order from the $diff_t$ values obtained for the other relations). As a result, the underlying quasi order is more frequently recovered. This is true particularly for smaller error probabilities. In addition, note that in the case of a large number of implications in the underlying quasi order, there are large differences of the sizes of the true and the neighboring relations in the selection set (due to transitivity). For instance, for nine items used in the simulation study, an underlying quasi order consisting of 64 implications has possible nearest neighbors which contain 58 or 72 implications, and the former even may not be included in the selection set. As a consequence, for an underlying large quasi order, missing the true relation already implies a large *dist* value.

6 Discussion

6.1 Summary

The original, corrected, and minimized corrected IITA algorithms (respectively, *diff* fit measures) have been proposed for building quasi orders from dichotomous data. Up until now, they have been treated descriptively, without examining theory underlying these procedures. In this paper, we have introduced maximum likelihood methodology, as a possible framework for investigating the methods theoretically. More precisely, the *diff* fit measures are seen to be maximum likelihood estimators for their corresponding population values (after some transformation using sample size). These estimators satisfy the desirable property of being asymptotically efficient, and hence they are asymptotically normal and unbiased, and consistent. In simulation studies, the three algorithms have been compared regarding finite sample consistency, and ranks and symmetric differences in population—not sample—quantities. The corrected and minimized corrected versions are seen to provide considerable improvements on the original algorithm.

6.2 Asymptotic variances, confidence intervals, and hypotheses testing

In a next step, we plan to derive population and estimated asymptotic variances of the maximum likelihood estimators $diff_t$. Consistent estimators for the population asymptotic variances can be obtained based on the expected and observed Fisher information matrices. The asymptotic variances can also be estimated using resampling methods such as the bootstrap. Large sample

normality with associated standard errors can be used to construct confidence intervals for the population values of and to test hypotheses about the *diff* coefficients. Confidence intervals allow for more informative estimation of the population values, compared to the maximum likelihood point estimates. This could be helpful in comparing *diff* values for different elements of a selection set. Moreover, one could, for instance, propose approximate significance tests for testing whether the population *diff* value for one quasi order is greater than the population value obtained for another (which is the crucial hypothesis to be tested when choosing among competing quasi orders). The quasi orders could, for example, be derived from different psychological theories.

6.3 Application of the paper's approach to other descriptive measures

The maximum likelihood approach to fit measures presented in this paper is a general one, in the sense that it can be applied to any, sufficiently smooth, coefficient for multinomial count data. After possibly performing necessary transformations to the coefficient, simply replace the relative cell frequencies by their corresponding population multinomial probabilities, and utilize the invariance property. In particular, this approach allows introducing maximum likelihood inference methodology for measures assessing the fit of general knowledge structures (not necessarily being a space or quasi ordinal).
We briefly exemplify that with the discrepancy index, *di*, by Kambouri, Koppen, Villano, & Falmagne (1994). The *di* coefficient (from the data to the knowledge structure) can be transformed and written as

$$di(D,\mathcal{K}) = \sum_{R\in 2^Q} \frac{m(R)}{m} d(R,\mathcal{K}),$$

where $d(R,\mathcal{K}) = \min_{K\in\mathcal{K}} d(R,K)$ and $d(R,K) = |(R \setminus K) \cup (K \setminus R)|$. The corresponding population *di* coefficient, obtained replacing $m(R)/m$ by $\rho(R)$, is $\sum_{R\in 2^Q} \rho(R) d(R,\mathcal{K})$. According to the invariance property, the sample *di* coefficient is the maximum likelihood estimator for this population value. Analogously as we have discussed for the *diff* coefficients, maximum likelihood inference methodology can be introduced for the *di* coefficient.
In other words, the approach of this paper is not restricted to quasi orders or quasi ordinal knowledge spaces, but can be applied to any, sufficiently smooth, measure that aggregates the multinomial cell counts according to an input rule for combining them, for instance, as determined by a quasi order or knowledge structure.

6.4 Résumé

The fit measures around in KST, whether they are formulated at the level of items or at the level of knowledge states, all aggregate the manifest multinomial cell counts into a single real number. This is why, uniformly, they can be based theoretically using the maximum likelihood approach. However, it is important to note that, in practice, the quality of the asymptotics has to be checked for finite sample sizes. In this paper, we have exemplified that with the consistency property for the *diff* coefficients, in simulation. For simpler aggregations (coefficients), closed form expressions for the finite sample moments of the estimators are possible of course.
Incorporating latent parameters into the formulations of the *diff* coefficients (or of other fit measures) is important. The manifest γ_{L} parameter in *diff* is used as an estimate of the latent response error probability. Instead, the expected numbers of counterexamples could be parameterized directly in terms of latent (e.g., careless error and lucky guess) parameters. Though the introduction of latencies may complicate theory and computation, it can provide for more realistic and interpretable results.
In future research, the present unitary approach using maximum likelihood should definitely be compared with other approaches to evaluating fit measures for KST models (e.g., Heller (2008), Schrepp (2007)).

References

Albert, D., Lukas, J. (Eds.), 1999. Knowledge Spaces: Theories, Empirical Research, and Applications. Lawrence Erlbaum Associates, Mahwah, NJ.

Bishop, Y.M.M., Fienberg, S.E., Holland, P.W., 1975. Discrete Multivariate Analysis: Theory and Practice. M.I.T. Press, Cambridge, MA.

Casella, G., Berger, R.L., 2002. Statistical Inference (2nd edition). Duxbury, Pacific Grove, CA.

Doignon, J.-P., Falmagne J.-Cl., 1999. Knowledge Spaces. Springer, Berlin.

Goodman, L.A., Kruskal W.H., 1979. Measures of Association for Cross Classifications. Springer, New York.

Heller, J., 2008. Empirical validation of knowledge structures: A resampling perspective. 39th Meeting of the European Mathematical Psychology Group, Graz, Austria, September 7-11.

Kambouri, M., Koppen, M., Villano, M., & Falmagne, J.-Cl., 1994. Knowledge assessment: Tapping human expertise by the QUERY routine. International Journal of Human-Computer Studies 40, 119-151.

R Development Core Team, 2006. R: A Language and Environment for Statistical Computing (ISBN 3-900051-07-0). R Foundation for Statistical Computing, Vienna, Austria.

Sargin, A., Ünlü, A., 2008. Inductive item tree analysis: Corrections, improvements, and comparisons. Manuscript submitted for publication.

Schrepp, M., 1999. On the empirical construction of implications between bivalued test items. Mathematical Social Sciences 38, 361-375.

Schrepp, M., 2003. A method for the analysis of hierarchical dependencies between items of a questionnaire. Methods of Psychological Research Online 19, 43-79.

Schrepp, M., 2007. On the evaluation of fit measures for quasi-orders. Mathematical Social Sciences 53, 196-208.

Witting, H., Müller-Funk, U., 1995. Mathematische Statistik II [Mathematical Statistics II]. Teubner, Stuttgart, Germany.

Zehna, P.W., 1966. Invariance of maximum likelihood estimators. The Annals of Mathematical Statistics 37, 744.

7 Tables

Table 1
Relative frequencies of 5000 data matrices (50 data matrices per one out of 100 quasi orders) satisfying $|\hat{\theta}_n - \theta| > \epsilon$; first, second, and third lines refer to the original, corrected, and minimized corrected IITA algorithms, respectively.

	Sample size			
	100	1000	10000	25000
ϵ				
0.01	0.0010	0	0	0
	0	0	0	0
	0	0	0	0
0.001	0.2402	0.0278	0	0
	0.0466	0	0	0
	0.0326	0	0	0
0.0001	0.9266	0.5636	0.4910	0.5240
	0.9540	0.2306	0.0158	0.0066
	0.9646	0.1878	0.0032	0.0002

Table 2
Average *dist*, *dist**, and *rk* values; first, second, and third lines refer to the original, corrected, and minimized corrected IITA algorithms, respectively.

	Summary statistic		
	dist	*dist**	*rk*
τ			
0.03	0.74	2.42	1.78
	3.10	1.72	1.60
	2.99	0.77	1.43
0.05	1.16	11.73	2.30
	2.76	2.23	1.68
	2.31	0.91	1.35
0.08	4.05	40.85	3.88
	3.72	2.17	1.95
	3.50	1.13	1.57
0.10	6.17	79.44	6.54
	3.59	2.89	2.35
	3.00	1.65	1.67
0.15	15.11	142.90	11.76
	3.62	6.56	3.18
	3.49	3.54	2.42
0.20	32.79	174.80	16.96
	4.56	14.76	4.79
	3.82	10.81	3.86

8 Figure caption

Fig. 1. Boxplots for the three IITA algorithms, within each of the sample sizes of the 50 computed sample $diff_t$ values. The three population $diff_t$ values are shown as horizontal lines in the plots.

9 Figure

Fig. 1