Analysis of Multinomial Response Data: a Measure for Evaluating Knowledge Structures

Ali Ünlü

Department of Psychology
University of Graz
Universitätsplatz 2/III
A-8010 Graz, Austria
(e-mail: ali.uenlue@uni-graz.at)

Abstract. Multinomial response data obtained from nominally and dichotomously scored test items in knowledge space theory are explained by knowledge structures. A central problem is the derivation of a "realistic" explanation, i.e., knowledge structure, representing the organization of "knowledge" in a domain and population of reference. In this regard, often, one is left with the problem of selecting among candidate competing explanations for the data. In this paper, we propose a measure for the selection among competing knowledge structures. The approach is illustrated with simulated data.
Keywords: Discrete multivariate response data, Qualitative test data analysis, Knowledge space theory, Selection measure, Simulation.

1 Knowledge space theory (KST)

This section reviews basic deterministic and probabilistic concepts of KST. For details, refer to [Doignon and Falmagne, 1999].

Definition 1 *A knowledge structure is a pair* $(Q, \mathcal{K})$*, with* Q *a non-empty, finite set, and* $\mathcal{K}$ *a family of subsets of* Q *containing at least the empty set* $\emptyset$ *and* Q*. The set* Q *is called the domain of the knowledge structure. The elements* $q \in Q$ *and* $K \in \mathcal{K}$ *are referred to as (test) items and (knowledge) states, respectively. We also say that* $\mathcal{K}$ *is a knowledge structure on* Q*.*

The general definition of a knowledge structure allows for infinite item sets as well. However, throughout this work, we assume that Q is finite.

The set Q is supposed to be a set of *dichotomous* items. In this paper, we interpret Q as a set of dichotomous questions/problems that can either be *solved* (coded as 1) or *not solved* (coded as 0). Here, "solved" and "not solved" stand for the observed responses of a subject (*manifest level*). This has to be distinguished from a subject's true, unobservable knowledge of the solution to an item (*latent level*). In the latter case, we say that the subject is *capable of mastering* (coded as 1) or *not capable of mastering* (coded as 0) the item. For a set X, let 2^X denote its *power-set*, i.e., the set of all subsets of X. Let $|X|$ stand for the *cardinality* (*size*) of X. The observed responses of a subject to the items in Q are represented by the subset $R \subset Q$ containing exactly the items that are solved by the subject. This subset R is called the *response pattern* of the subject. Similarly, the true latent state of knowledge of a subject with respect to the items in Q is represented by the subset $K \subset Q$ containing exactly the items the subject is capable of mastering. This subset K is called the *knowledge state* of the subject. Given a knowledge structure $\mathcal{K}$, we assume that the only states of knowledge possible are the ones in $\mathcal{K}$. In this sense, $\mathcal{K}$ captures the organization of knowledge in the domain and population of reference. Idealized, if no response errors, i.e., careless errors and lucky guesses, would be committed, the only response patterns possible would be the knowledge states in $\mathcal{K}$.

Let $\mathbb{N}$ stand for the set of natural numbers (without 0). We fix a population of reference, and examinees are drawn from this population randomly. Let the sample size be $N \in \mathbb{N}$. The data is constituted by the observed absolute counts $N(R) \in \mathbb{N}_0 := \mathbb{N} \cup \{0\}$ of response patterns $R \in 2^Q$. The data, $\mathbf{x} = (N(R))_{R \in 2^Q}$, are assumed to the realization of a random vector $\mathbf{X} = (X_R)_{R \in 2^Q}$, which is distributed *multinomially* over 2^Q. That is,

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

Here, $\rho(R) \in [0,1]$ for any $R \in 2^Q$, $\sum_{R \in 2^Q} \rho(R) = 1$, and $N(R) \in \mathbb{N}_0$ with $0 \leq N(R) \leq N$ for any $R \in 2^Q$, $\sum_{R \in 2^Q} N(R) = N$.

Let the maximum probability of occurence be denoted by $\rho(R_m)$, i.e.,

$$\rho(R_m) = \max_{R \in 2^Q} \rho(R),$$

for some appropriate response pattern $R_m \in 2^Q$.

Maximum likelihood estimates (briefly, MLEs) for the population probabilities $\rho(R)$ $(R \in 2^Q)$ are $\widehat{\rho(R)} = N(R)/N$. The MLE for $\rho(R_m)$ is $\widehat{\rho(R_m)} = N(R'_m)/N$, where $N(R'_m)$ denotes the maximum absolute count $N(R'_m) = \max_{R \in 2^Q} N(R)$, for some appropriate response pattern $R'_m \in 2^Q$.

We will simulate multinomial response data in accordance with a *basic local independence model*.

Definition 2 *A quadruple $(Q, \mathcal{K}, p, r)$ is called a basic local independence model (BLIM) iff*

1 *$(Q, \mathcal{K})$ is a knowledge structure;*

2 *p is a probability distribution on $\mathcal{K}$, i.e., $p : \mathcal{K} \to [0,1], K \mapsto p(K)$, with $p(K) \geq 0$ 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 is a function $r : 2^Q \times \mathcal{K} \to [0,1]$, $(R, K) \mapsto r(R, K)$, 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.,*

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

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

A probability distribution p on $\mathcal{K}$ (point **2**) is interpreted as follows. To each knowledge state $K \in \mathcal{K}$ is attached a probability $p(K) \in [0,1]$ measuring the likelihood that a randomly sampled subject is in state K. Further, any randomly sampled subject is necessarily in exactly one of the states of $\mathcal{K}$. A response function r (point **3**) is interpreted as follows. For $R \in 2^Q$ and $K \in \mathcal{K}$, $r(R,K) \in [0,1]$ specifies the conditional probability of response pattern R for an examinee in state K. Given the probability distributions p on $\mathcal{K}$ and $r(\,.\,,K)$ on 2^Q ($K \in \mathcal{K}$), a BLIM takes into account the two ways in which probabilities must supplement deterministic knowledge structures. For one, knowledge states will occur with different proportions in the population of reference. For another, response errors (careless errors and lucky guesses) will render impossible the a-priori specification of the observable responses of a subject, given her/his knowledge state. The condition of *local independence* (point **4**) states that the item responses of an examinee are assumed to be independent, given the knowledge state of the examinee, and the response error probabilities $\beta_q, \eta_q \in [0,1[$ ($q \in Q$) are attached to the items and do not vary with the knowledge states.

The BLIM is a *multinomial* probability model.

Corollary 1 *Given a BLIM, the occurence probabilities of response patterns are parameterized as*

$$\rho(R) = \sum_{K \in \mathcal{K}} \left\{ \left[\prod_{q \in K \setminus R} \beta_q \right] \cdot \left[\prod_{q \in K \cap R} (1-\beta_q) \right] \cdot \left[\prod_{q \in R \setminus K} \eta_q \right] \cdot \left[\prod_{q \in Q \setminus (R \cup K)} (1-\eta_q) \right] \right\} p(K).$$

□

2 Measure $\boldsymbol{\kappa}$

In this section, we propose a measure, κ, for the selection among competing explanations, i.e., knowledge structures, for the multinomial response data. For details, refer to [Ünlü, 2004].

2.1 Prediction paradigm

The derivation of κ heavily rests on the following *prediction paradigm.*

The *prediction problem* considered is this. An individual is chosen randomly from the population of reference, and we are asked to guess his/her response pattern, given, either

(no info). no further information (than the multinomial distribution), or
(info). the knowledge structure $\mathcal{K}$ assumed to underlie the responses of the individual.

The *prediction strategies* in both cases are as follows. In the "no info" case, we *optimally* guess some response pattern $R_m \in 2^Q$, which has the largest probability of occurence $\rho(R_m) = \max_{R\in 2^Q} \rho(R)$. In the "info" case, we *proportionally* guess the knowledge states with their probabilities of occurence. That is, if $\mathcal{K} = \{K_1, K_2, \ldots, K_{|\mathcal{K}|}\}$, we guess K_1 with probability $\rho(K_1)$, K_2 with probability $\rho(K_2)$, ..., $K_{|\mathcal{K}|}$ with probability $\rho(K_{|\mathcal{K}|})$. Since these probabilities may not add up to one, in general, there is a non-vanishing *residual* probability $\{1 - \sum_{K\in\mathcal{K}} \rho(K)\} > 0$. Thus, in order to complete the prediction strategy, we abstain from any guessing with probability $1-\sum_{K\in\mathcal{K}} \rho(K)$, and, in the sequel, view this as a prediction error.

The probabilities of a *prediction error* in both cases are as follows. In the "no info" case, the probability is $1 - \rho(R_m)$, and, in the "info" case, it is $1 - \sum_{K\in\mathcal{K}} \rho(K)^2$. Of course, the probabilities of a *prediction success* are $\rho(R_m)$ and $\sum_{K\in\mathcal{K}} \rho(K)^2$, respectively.

2.2 First constituent of κ: measure of fit

The measure κ consists of two constituents. The first constituent of κ captures the (descriptive) *fit* of a knowledge structure $\mathcal{K}$ to the response data. This constituent is derived based on the method of *proportional reduction in predictive error* (PRPE)—the method of PRPE was introduced originally by [Guttman, 1941], and it was applied systematically in the series of papers by [Goodman and Kruskal, 1954, 1959, 1963, 1972]. The general probability formula of the method of PRPE quantifies the *predictive utility*, $PU(info)$, of given information. Informally,

$$PU(\mathit{info}) := \frac{\text{Prob. of error (no info)} - \text{Prob. of error (info)}}{\text{Prob. of error (no info)}}.$$

Inserting the previous prediction error probabilities into the PRPE formula, we obtain the population analogue of the first constituent, m_1, of κ.

Definition 3 *Let $\rho(R_m) \neq 1$. The measure m_1 is defined as*

$$m_1 := \frac{\big(1-\rho(R_m)\big) - \big(1-\sum_{K\in\mathcal{K}} \rho(K)^2\big)}{1-\rho(R_m)} = \frac{\sum_{K\in\mathcal{K}} \rho(K)^2 - \rho(R_m)}{1-\rho(R_m)}.$$

In the sequel, we assume that $\rho(R_m) \neq 1$! Inserting MLEs, we obtain the MLE, $\widehat{m_1}$, for m_1 (We assume that $1 - N(R'_m)/N \neq 0$!):

$$\widehat{m_1} = \frac{\sum_{K \in \mathcal{K}} N(K)^2 - N \cdot N(R'_m)}{N^2 - N \cdot N(R'_m)}.$$

2.3 Second constituent of κ: measure of size

The second constituent of κ captures the *size* of a knowledge structure $\mathcal{K}$. For the definition of it, we need the concept of a *truncation* of $\mathcal{K}$.

Definition 4 *Let $M \in \mathbb{N}$ be a truncation constant. An M-truncation of $\mathcal{K}$ is any subset, $\mathcal{K}_{M\text{-}trunc}$, of $\mathcal{K}$ which is derived in the following way.*

1. *Order the knowledge states $K \in \mathcal{K}$ according to their probabilities of occurence $\rho(K)$, say, from left to right, ascending with smaller ρ values to larger ones. Knowledge states with equal probabilities of occurence are ordered arbitrarily.*
2. *Starting with the foremost right knowledge state, i.e., a knowledge state with largest probability of occurence, take the first $\min(|\mathcal{K}|, M)$ knowledge states, descending from right to left. The set of these knowledge states is $\mathcal{K}_{M\text{-}trunc}$.*

The definition of the second constituent, m_2, of κ is this.

Definition 5 *Let $\sum_{K \in \mathcal{K}} \rho(K) \neq 0$. Let $M \in \mathbb{N}$ be a truncation constant, and let $\mathcal{K}_{M\text{-}trunc}$ denote an M-truncation. The measure m_2 is defined as*[1]

$$m_2 := \frac{\sum_{K \in \mathcal{K}} \rho(K)^2}{\sum_{K \in \mathcal{K}_{M\text{-}trunc}} \rho(K)^2}.$$

In the sequel, we assume that $\sum_{K \in \mathcal{K}} \rho(K) \neq 0$ for any knowledge structure $\mathcal{K}$. Inserting MLEs, we obtain the MLE, $\widehat{m_2}$, for m_2 (We assume that $\sum_{K \in \mathcal{K}} N(K) \neq 0$!):

$$\widehat{m_2} = \frac{\sum_{K \in \mathcal{K}} N(K)^2}{\sum_{K \in \mathcal{K}_{\widehat{\text{M-trunc}}}} N(K)^2},$$

where $\mathcal{K}_{\widehat{\text{M-trunc}}}$ is defined analogously as in Definition 4, where we have to replace $\rho(K)$ with its MLE $N(K)/N$ for any $K \in \mathcal{K}$.

[1] m_2 is invariant with respect to the choice of a particular M-truncation.

2.4 κ: size trading-off fit measure

The measure κ is (more or less) the product of m_1 and m_2.

Definition 6 *Let $M \in \mathbb{N}$ be a truncation constant, and let $C \in [0, 0.01]$ be a small, fixed non-negative correction constant.*[2] *The measure κ is defined as*

$$\kappa := m_2 \cdot (m_1 - C).$$

The MLE for κ is $\widehat{\kappa} := \widehat{m_2} \cdot (\widehat{m_1} - C)$.

The measure κ may be interpreted as a *performance measure* for the evaluation of knowledge structures. The two (performance) criteria being merged and traded-off are "(descriptive) fit" and "(structure) size", respectively measured by its constituents m_1 and m_2. The *decision rule* important for applications of κ is this. *The greater the value of κ is, the "better" a knowledge structure "performs" with respect to a trade-off of the criteria.* The (unknown) ordering of the population κ values is "estimated" by the ordering of the corresponding MLEs.

2.5 Model selection and truncation constant

Finally, we describe a special choice for the truncation constant in the context of model selection among competing knowledge structures $\mathcal{K}_1, \mathcal{K}_2, \ldots, \mathcal{K}_n$ ($n \in \mathbb{N}$, $n \geq 2$) on (same) domain Q.

Definition 7 *Let $v_i := |\{K \in \mathcal{K}_i : \rho(K) \neq 0\}|$ be the match of candidate model $\mathcal{K}_i$ ($1 \leq i \leq n$). Let $\mathbf{v} := (v_1, v_2, \ldots, v_n)^T \in \mathbb{N}^n$ be the match vector. The (empirical) median of the matches $v_i \in \mathbb{N}$ ($1 \leq i \leq n$) is denoted by median($\mathbf{v}$) and called the median match of the competing models. Formally,*

$$median(\mathbf{v}) := \begin{cases} v_{(\frac{n+1}{2})} & : \ odd\ n \\ v_{(\frac{n}{2})} & : \ even\ n, \end{cases}$$

where $v_{(1)}, v_{(2)}, \ldots, v_{(n)}$ with $v_{(1)} \leq v_{(2)} \leq \cdots \leq v_{(n)}$ is the ordered list of matches v_i ($1 \leq i \leq n$).

The special truncation constant, M_s, is this.

Definition 8 *The special truncation constant M_s is defined as*[3]

$$M_s := \min\left([2^{|Q|/2}], median(\mathbf{v})\right).$$

[2] C is introduced to compensate for a zero value of m_1.

[3] The meaning of term $2^{|Q|/2}$ is clarified in the context of knowledge assessment procedures (for details, see [Ünlü, 2004]). For any real $x \geq 0$, $[x]$ denotes the entier of x, i.e., the integer $I \in \mathbb{N} \cup \{0\}$ with $I \leq x < I + 1$.

3 Simulation example

In this section, we apply κ to data simulated in accordance with a specific BLIM. For details (including software), refer to [Ünlü, 2004].

We consider the knowledge structure

$$\mathcal{H} := \Big\{\emptyset, \{a\}, \{b\}, \{a,b\}, \{a,b,c\}, \{a,b,d\}, \{a,b,c,d\}, \{a,b,c,e\}, Q\Big\}$$

on domain $Q := \{a,b,c,d,e\}$. We suppose that the knowledge states of $\mathcal{H}$ occur in a population of reference with the probabilities

$$\begin{aligned}
p(\emptyset) &:= 0.04,\\
p(\{a\}) &:= 0.10,\\
p(\{b\}) &:= 0.06,\\
p(\{a,b\}) &:= 0.12,\\
p(\{a,b,c\}) &:= 0.11,\\
p(\{a,b,d\}) &:= 0.07,\\
p(\{a,b,c,d\}) &:= 0.13,\\
p(\{a,b,c,e\}) &:= 0.18,\\
p(Q) &:= 0.19.
\end{aligned}$$

Let the careless error and lucky guess probabilities β_q and η_q at items $q \in Q$, respectively, be specified as

$$\begin{aligned}
\beta_a &:= 0.16,\ \eta_a := 0.04,\\
\beta_b &:= 0.18,\ \eta_b := 0.10,\\
\beta_c &:= 0.20,\ \eta_c := 0.01,\\
\beta_d &:= 0.14,\ \eta_d := 0.02,\\
\beta_e &:= 0.24,\ \eta_e := 0.05.
\end{aligned}$$

Based on this BLIM, we simulated a binary (of type 0/1) $1\,200 \times 5$ data matrix representing the response patterns for 1 200 fictitious subjects. The collection of competing models (knowledge structures) for model selection was obtained from the multinomial response data data-analytically, based on a modified version of the *Item Tree Analysis* (ITA; see [Leeuwe, 1974]) described in [Ünlü, 2004]. A modified ITA of the BLIM data resulted in a collection of fifteen knowledge structures, which contained the true knowledge structure $\mathcal{H}$ underlying the data.

From this collection, we selected an optimal model based on maximum κ. Table 1 lists the values of κ (for $M := M_s$, and $C := 0.01$) for the fifteen competing knowledge structures. In Table 1, models are labeled by their respective *tolerance levels* $0 \leq L \leq 1200$ of the modified ITA, and L_κ denotes the optimal (maximum κ) solution. The true model is labeled by "(true)".

L	κ
0–58	−0.098487
59–62	−0.098591
63–71	−0.098672
72–77	−0.098807
78–88	−0.098880
89–95	−0.098931
96–100	−0.099029
101–150 (true)	−0.099040
151–191	−0.098871
L_κ = 192–213	−0.097610
214–236	−0.098913
237–239	−0.102439
240–285	−0.108678
286–394	−0.118036
395–1 200	−0.133919

Table 1. κ (for $M := M_s$, and $C := 0.01$)

Measure κ assumed its maximum value at tolerance range $L_\kappa = 192\text{–}213$, i.e., for the candidate knowledge structure $\mathcal{K}_{192-213}$,

$$\mathcal{K}_{192-213} := \left\{\emptyset, \{a\}, \{a,b\}, \{a,b,c\}, \{a,b,c,d\}, \{a,b,c,e\}, Q\right\}.$$

Compared to the true model $\mathcal{K}_{101-150} = \mathcal{H}$, this "best" solution was quite acceptable. In $\mathcal{H}$, the subsets $\{b\}$ and $\{a,b,d\}$ were knowledge states, whereas, in $\mathcal{K}_{192-213}$, they were not. In all other respects, both the models were identical. We had $|\mathcal{H}| = 9$ versus $|\mathcal{K}_{192-213}| = 7$ ($\mathcal{K}_{192-213} \subset \mathcal{H}$).

Acknowledgements

This work was supported by the Austrian Science Fund (FWF) through the project grant P17071-N04 to Dietrich Albert.

References

[Doignon and Falmagne, 1999]J.-P. Doignon and J.-Cl. Falmagne. *Knowledge Spaces.* Springer, Berlin, 1999.

[Goodman and Kruskal, 1954]L.A. Goodman and W.H. Kruskal. Measures of association for cross classifications. *Journal of the American Statistical Association*, pages 732–764, 1954.

[Goodman and Kruskal, 1959]L.A. Goodman and W.H. Kruskal. Measures of association for cross classifications, II: Further discussion and references. *Journal of the American Statistical Association*, pages 123–163, 1959.

[Goodman and Kruskal, 1963]L.A. Goodman and W.H. Kruskal. Measures of association for cross classifications, III: Approximate sampling theory. *Journal of the American Statistical Association*, pages 310–364, 1963.

[Goodman and Kruskal, 1972]L.A. Goodman and W.H. Kruskal. Measures of association for cross classifications, IV: Simplification of asymptotic variances. *Journal of the American Statistical Association*, pages 415–421, 1972.

[Guttman, 1941]L. Guttman. An outline of the statistical theory of prediction. In P. Horst et al., editors, *The Prediction of Personal Adjustment*, volume 48 of *Bulletin*, pages 253–318. Social Science Research Council, New York, 1941.

[Leeuwe, 1974]J.F.J. van Leeuwe. Item tree analysis. *Nederlands Tijdschrift voor de Psychologie*, pages 475–484, 1974.

[Ünlü, 2004]A. Ünlü. *The Correlational Agreement Coefficient CA and an Alternative* κ. PhD thesis, Graz University of Technology, Graz, Austria, 2004.