3mc2008 – Proceedings ©

Representations of ordinality and substantive research hypotheses in cross-country comparisons based on survey data

A. Jan Kutylowski
andrzejk@ulrik.uio.no
akutylow@ifispan.waw.pl

2008/05/31

Abstract

It is typical for multinational survey programs, such as International Social Survey Program (ISSP) or European Social Survey to include in questionnaires selected questions with ordinal categories. In special case such questions purport, sometimes tacitly, to related to a common underlying concept (such as e.g. trust) or hidden (latent) variables (such as national pride).

The intended presentation will systematize and integrate a number of ideas, dispersed but fairly well established in statistical methodology, for the purpose of representing the ordinality of the information inherent in such questions or in structures described by such questions. A number of frameworks will be considered.

First, graphical displays of ordinal variables will be briefly reviewed, with emphasis on *sequential* concept of ordinality and the efficiency of the latter in exploratory, cross-country comparisons.

Second, the idea of multiple indicators relating, in an non-overlapping manner to single hidden variable will be discussed. A class of semi-parametric graphical models, selected by Monte-Carlo methods, will be applied for the purpose of discovering structure of association among the indicators. Emphasis will be put again on exploratory uses of the method for the purpose of formulation of substantive research hypotheses across country boundaries. It will be further argued that the method elucidates the steps in (often uncritical) recourse to the prediction of hidden variable(s) by identifying subsets of variables whose associations should be explained by prediction of a hidden variable.

Third, the idea of nonparametric hidden variable model will be briefly outlined, with application to validation of the prediction of a hidden variable ("scale") in a flexible manner allowing partial evaluation of ordinality of an indicator w.r.t the hidden variable. It will also be shown that under mild assumptions such modelling is capable of rendering of categorical ordinal hidden variables as revealed by the data.

Applications will use selected data from ISSP and European Social Surveys. The level of the presentation will emphasize immediate practical uses of the proposed methods, and their advantages over inefficient procedures considered "standard" in substantive literature published in social science fields.

This is a draft version of the presentation. I plan at least two updates before the beginning of the conference. Please do not quote or cite without explicit permission from the author.

1 Introduction

Cross-country comparisons became a focus of social scientists considerable time ago. The advent of data bases in political science with state or country as a unit can be seen as an early minifestation of such interest in politology (also known as political science). It has been however relatively recently, that cross-country surveys, as opposed to other types of data, became *systematically* and *comprehensively* carried out, with the ensuing availablity od their data not only for academic social scientist but also for "the man on the street", including the student. Today comprehensive data sets from research programs such as ISSP, ESS, EVS or WVS[1] are available without restrictions, which offers new analytical opportunities.

This paper will focus on issues relate to the treatment of *ordinal* variables in such surveys. Ordinal variables are present elsewhere, where they are at least as important as in cross-country surveys. However, in the latter surveys they are used to a greater extent for the purpose of cross-country comparisions, which leads to a class of specific challenges related to the interpretation of ordinality.

2 Unirespone and multiresponse framework: two examples

Evans and Kelley (2001) discuss *national pride* in a comparative context with the *notion* of national pride as *specified* via inventory (a set of items), each with four ordinal categories, in the ISSP module "National Pride" of 1995 which encompass 24 countries. The items in question are treated as if the *category labels*, $1, 2, 3, 4$ of categories "not proud at all" (1), "not very proud" (2), "proud" (3) and "very proud" (4), were *values* of a hypothetical metrical variable, y^* say.

The original categorical variable, say Y is *multivariate* with 3 dimensions, since knowing probabilities of occurence of any three categories specifies the fourth probability (all probablities sum to 1 by definition]. The aforementioned authors *scale* the labels so that the highest category is allocated value 100 and the lowest category value 0, the intermediate categories being scored at equal intervals in between. From quantitative point of view this corresponds to a an expression, in percentage terms, of analogous scaling with the largest value equal to 1. Given this approach, whose implications (dimensionality reduction) are not discussed, it is straighforward to calculate mean of a single indicator for the countries of interest, group the countries into presumably meaningful classes (such as USA, other English speaking, Germanic, Nordic, "other [societies] with market [economies]" and "Eastern Europe") for the purpose of detailed comparisons. Includion of other indicators is straighforward, as they are seen as separate at the level of formal analysis. Interestingly, the authors do not calculate the variances for the pseudo-metrical indicators y_i^*, stating that they have "single-peaked, well behaved distributions well represented by their means (2001:313). Following this pseudo-metrical approach they calculate Pearson correlation coefficients for various indicators, means adjusted with respect to factors affecting the responses, as well as effects of factors, such as age, gender, education on the mean scores of given aspect of nation's pride. In such analysis possible differences in variances of pseudo-metrical response variables are not accounted for, it being assumed that the confitions of the *standard* Normal linear model hold.

Given the relative familiarity with and the elegance of the tools availale once pseudo-metrical variable is constructed from ordinal indicators, one is tempted to ask why was the ordinal indicator collected at all, if the ordinality in such analyses is not taken into account. If a method existed that would allow direct record of the metrical variable it should have been used instead. One could argue that such a method does not exist, so that there is a gap between the level of data collection

[1] I leave it to the reader to decipher those abbreviations, if necessary.

(or observation) and the level of data analysis. If we cannot analyse what we observe, then what is the point making the observation?

For the purpose of simplifying and systematizing terminology, category of a categorical variable will be interpreted as the sought *reaction* of the respondent to the question, whereas the variable representing those reactions will be interpreted as response. An inventory of question creates a multiresponse framework.

Smith and Jarkko (2001) present an analysis of comparative national pride based on same survey that takes into account 10 specified indicators of the pride in given domains, such as "the way democracy works", "history" or "fair and equal treatment of all groups in society". The authors assume that the unscaled category labels are values of the respective categorical variables, which are thus treated as metrical. They are interested in compound assessments and comparisons of nations. The first they obtain for each respondent in a given country as

$$y^* = \sum_{v=1}^{U} y_u^*. \tag{2.1}$$

Averaging over respondents enables thus, among others, ranking of the countries. Similarly, they consider "general" national pride as composed of five indicators with labels ranging from 1 to 5: applicationof same procedure results in respective second index of pride. Further analysis relies on Pearson correlation coefficients and systematic comparisons of classes of countries with respect to both indices.

Whereas first authors focussed on *uniresponse* framework, with only one variables in focus most of the time, the second authors were explicitly interested in *multiresponse* framework, where several response variables were conceived as informing about a single, underyling *hidden* (or *latent*) variable. Such variable can be interpreted as a *factor*, that is either *metrical* (quantitative) or categorical. Considering that the indicators were ordinal, on would expect that the categorical factor would be at least ordinal.

One can generally say that not only location, but also *dispersion* and *shape* will be of interest in cross-country analyses. It should be noted that the original observed variable is multicategory variable, which implies that when it is represented by a random variable, this will be a vector variable that is multivariate (multinomially distributed) with probabilities $\{\pi_u : 1 \leq u \leq U\}$. There is no obvious notion of dispersion and shape for such variables.

In the following I shall focus on some conceptual distinctions, and especially on some visual tools, that are connected to statistical models for categorical responses

3 Uniresponse framework

3.1 Bar charts

Figure 1 represents self-assessment of the location in one of 10 ordinal strata in four countries: Germany (divided into West and East Germany, the latter formerly under communist rule), United States, Japan and Russia. Bar charts for those countries illustrate differences in disparsion and shape of that categorical variable. Only some of the features will be mentioned:

- Dispersion of that distribution in Russia is much smaller than in other countries. There are also differences in skewness, and the most probable category "6" has very high probability of choice. There is clear skeweness to the right for West Germany, USA and Japan, but not for East Germany.

- The visual instrument shown here does rely on the assumption that categories are equidistant. As we see, this assumption does not preclude qualitative observations concerning dispersion and shape of the distribution.

One can smooth the distributions by employing line plot: this makes between-countries differences more visible (Figure 2).

Figure 1 about here: see consequtive pages with graphics following the end of the draft.

Figure 1: Smoothed distributions of self-assessed location in society in strata from 1 (lowest) to 10 (highest).

3.2 Cumulative probabilities

Alternative display of univariate categorical responses can be based on cumulative probabilities with accumulation corresponding to the ordinal sequence beginning at the lowest category. This can be easily implemented by declaring the variable to be metrical and choosing e.g. cumulative option in bar charts or, if not available, in histogram. Further discussion of this tool does not seem to be necessary.

Figure 2 about here: see consequtive pages with graphics following the end of the draft.

Figure 2: Smoothed distributions of self-assessed location in society in strata from 1 (lowest) to 10 (highest).

3.3 Sequential probabilities

The response Y is multinomially distributed with category probabilities π_u, $\sum \pi_u = 1$ so that dimensionality of the variable is $Q = U - 1$.

The cumulative approach mentioned in the above uses ordinal components (as function of category probabilities) τ that are defined as cumulative probabilities

$$\tau_u = \kappa_u \equiv P(Y \leq Y_u) = \sum_{v=1}^{u} \pi_u.$$

Cumulative response mechanism, while appealing, is not uniquely suitable for ordinal response data. Alternative *sequential* response mechanism can be motivated by reference to the behavior of respondents selecting levels of integer valued discrete response, e.g. number of children in a family. Here the selection will typically be sequential: beginning with no children, the couple can decide to have the second child only after it already has the first one, and analogously for the third child, *etc*.

In upward sequential selection a respondent begins with Y_1 and either reaches Y_2 or not: if Y_2 is reached the respondent may, or may not, reach Y_3, etc. Thus the Y_u'th alternative cannot be "agreed with" without "agreeing with" all the alternatives $Y_{u'}, u' < u$ beforehand: faced with Y_u respondent accepts it or rejects it. Each category is judged successively against all higher categories until some category is finally accepted. Conditionally on all preceding steps (or transitions) taken, each subsequent step involves a binary decision which is independent of precedent selections. For upward transitions, beginning with the lowest category, and with U=4 response components are the transition probabilities $\{\tau_u\}$: $P(Y_1; Y \geq Y_1) = \pi_1 \equiv \tau_1$, $P(Y_2; Y \geq Y_2) \equiv \tau_2$, $P(Y_3; Y \geq Y_3) \equiv \tau_3$, and $P(Y_4; Y \geq Y_4) = \pi_4 \equiv \tau_4$. Here τ_4 may be omitted without loss of information.

Upward sequential selection defines response components

$$\tau_u = \pi_u/(1 - \kappa_{u-1}), \qquad 1 \leq u < U.$$

It should be emphasized that variables corresponding to the sequential components are (asymptotically) independent and can be interpreted independently of each other, unless one would want to impose interpretational constraints across transitions.

Figure 3 shows sequential components for self-assessed membershiip in ordinal societal strata for the four countries previously concerned.

Figure 3 about here: see consequtive pages with graphics following the end of the draft.

Figure 3: Sequential components (upward conditional probabilities) of ordinal response variable “self-assessed location in society in strata from 1 (lowest) to 10 (highest)” By definition the last component must be equal to 1 and is thus redundant.

The interpretation of cross-country differences concerns itself in this case with substantively different aspects of the responses. One striking difference concerns the conditional probability in West Germany of “entering” stratum 9 given that one has “reached” stratum 8: this probability is considerably higher then in all other countries concerned, which points to specific forms of inclusiveness in that country, not yet present in the (former) East Germany, where all other sequential components show fairily similar values.

A special case of sequential modelling arises when the linearizing transformation of sequential probabilities is the logit function,

$$\mathrm{lgt}(\pi) \equiv \lambda = \ln[\pi/(1 - \pi)].$$

Then

$$\mathrm{lgt}(\tau_u) = \ln[\pi_u/(1 - \kappa_u)] = \psi_u. \tag{3.1}$$

It is important to note that taking $\mathrm{lgt}(\tau_u)$ with the minus signs gives

$$\ln \rho_u \equiv -\,\mathrm{lgt}(\tau_u) = \ln[(1 - \kappa_u)/\pi_u] = P(Y > Y_u; Y \leq Y_u), \tag{3.2}$$

where ρ_u are odds of continuing upward selection given that respondent reached the current step. These odds, typically defined in terms of expected multinomial frequencies, have been called *continuation ratios*. It follows that calculation of continuation ratios leads to specialized form of cross-country display shown in Figure 3. Further specialization of such displays follows if one desides to apply logit transformation to the continuation ratios.

4 Multiresponse framework

4.1 Data

Attitudinal data in cross-country surveys typically involve an inventory (a set of indicators) consisting of “ratings” that are (i) *homogeneous* (or identical) for each response (i.e. identically defined ordinal categories for each item involved), and that (ii) aim at eliciting supposedly equidistant gradings of reactions, i.e. response categories. One could say that the latter feature distinguishes *Likert inventories* from other instances of ordinal multiresponse data. Although (i) is often violated by sociological inventories, we shall mainly focus, for brevity of notation but without loss in generality, on homogeneous ratings. On many occasions the term “inventory” is called “scale”, as if the *indicand* to which the indicators (items) presumably relate, was the same as the set of indicators. Such imprecision is likely to lead to confusion, as when the “score” obtained in intelligence test is mistaken for the intelligence itself.

Occupational inventory data comprise a cross-classification of k ordinal categorical response variables, where item categories (here called *reactions*) are indexed in increasing order by $u(i)$, $1 \leq u(i) \leq U(i)$. For homogeneous items $U(i) = U$ for all i. Combined reactions (response categories) for a given respondent and a given inventory will be called *reaction profile*, which can be indexed in a specified manner by u, $1 \leq u \leq U$. Since the total number of reaction profiles is typically much larger than N, the data are highly sparse.

An example of multiresponse framework is provided by a set of questions from the ISSP modules on national identity concerning *natiocentrism* (in analogy to ethnocentrism) in terms of graded support (from strong disagreement to strong agreement) for five types of politicies. The policies concern (A) protective restriction of imports (*ImportLim*), (B) right of international bodies to impose solutions concerning environmental pollution (*IntBodies*), (C) implementation of politices following own interests, even if this leads to conflict (*OwnInterest)*, (D) not allowing foreigners to own land (*NoLand*), and also (E) preferential treatment of domestic TV films and programs (*OwnTV*). This is an example of “short” inventory, and thus relatively less sparse data. Shorter inventories appear to be in use, but considerably longer inventories also are prevalent. An example of the latter concerns are untypical,

4.2 Semiparametric approach in terms of coefficients of association

With two categorical variables A and B one can test the hypothesis of independence using e.g. “standard´´ test statistic asymptotically distributed as χ^2. With three categorical variables marginal independence of A and B may however coincide with conditional dependence (association) of A and B given C. Conditional association may take peculiar, nonintuitive forms, e.g.it may be concentrated in certain strata of C only.[2] It adds to the complexity of the problem that this is also true when there are more variables to potentially condition on. Additionally,

1. when the dimensionality of variables to condition on increases, data becames sparse.
2. With sparse data well-known test statistics applicable under hypothesis of independence (no association) are poorly approximated by their asymptotic distributions.
3. Additionally, these statistics do not take ordinality of the variables into account, which makes “standard” tests weak, i.e. they lack power against “unspecified alternatives” to independence.

[2] Some such forms are called paradoxes, e.g. Simpson’s paradox or, much earlier and therefore lesser known Yule’s paradox. Their discussion is outside the scope of this presentation.

The first postulate deriving from the above observations states, that it would be preferable to reformulate "plain" zero hypothesis of independence in terms of a coefficient of ordinal association, where specified value of that coefficient (typically 0) correspondes in a sense with the hypothesis of independence. When this hypothesis is rejected (with greater power than in the "standard" situation), should be thus able to estimate confidence intervals for the coefficient.

The second postulate states, that suitable coefficient should be tractable, i.e. it should have a partial version allowing tests of conditional indepence (no conditional association) and thus, *ipso facto*, estimation of the strength of conditional associations.

The well known $gamma$ coefficient of *ordinal* association, proposed by Goodman and Kruskal has a partial version developed by Davis (1967) and convenient asymptotic properties enabling calculation of confidence intervals and asymptotic tests. As such it is especially suited for investigating association structures of ordinal inventories.

Unreality of asymptotic properties of $gamma$ and $partialgamma$ can be dealt with by computationally intensive *exact* or less computationally intensive simulated (or Monte Carlo) exact tests, that can be done adaptively, in the sense of sequential or repeated Monte Carlo (MC) tests with desired level of accuracy. All these components, taken altogether, provide a viable semi-parametric toolbox for investigating association structures across countries. These can be represented either as

1. matrices indicating partial (supplemented by marginal) independencies (together with resulting test statistics) or, more comprehensively, or
2. matrices containig estimates of partial γ coefficients, with resulting test statistics based on asyptotic (not recommended), exact (computationally problematic) or simulated exact (MC) test statistics.

Illustration will be provided
of estimation and testing of partial γ coefficients for the five-variable example displayed in the next figure.

Qualitative differences in patterns of conditional independencies or in quantitative differences in association patterns (represented by estimates of γ) are suitable for revealing cross-country differences and similarities. However, most statistical packages lack essential components of the toolbox involved. One exception is the availability of extraneous Stata program for estimation of partial gamma coefficients developed by Kreiner and Lauritzen (to be specified)

4.3 Graphical models of association patterns

The following figure displays a graphical model obtained by model selection starting from the baseline "full" model, where each variable is connected to each other. Data from 2003 ISSP module for Japan were used.

Figure 4 about here: see consequtive pages with graphics following the end of the draft.

Figure 4: Selected graphical model following a selction procedure starting with full model (all variables connected with each other).

The most striking feature of the resulting graph concerns pattern of conditional independencies. Most notably *OnwnInterest* (C) is conditionally independent of *IntBodies* (B) given *NoLand* (D). Opinion concerning *NoLand* separates two subgraphs ECD and DBA. An attempt to modify the graph, and thus its the association structure would be most successful *if* the strength of (D) were modified.

Despite its relatively small size, this inventory of notiocentrism clearly illustrates the "curse of dimentionality" that makes statistical inference difficult and in extreme cases impossible. Five variables each with five categories requires at least "3,125" observations to fill all cells in the joint contingency table representing joint distribution of responses. In most cases observations corresponding to about one third of that number constitute total sample.

However, it follows from general properties of graphical models, that testing of conditional associations may be conducted using only certain subgraphs, which makes inferences possible.[3]

4.4 Nonparametric modelling of association patterns with a hidden variable

The rationale for introducing hidden variable into the model typically presupposes existence of a complex association pattern that can be usefully summarised by conditional independence of all observed responses given the hidden variable. Next Figure shows an instance of such pattern. The data concern opinions as to the importance of a number of aspects connected to national identity of Japanese (ISSP 2003). Attempts at simplifying the model by removing at least one of the vertices (and thus introducing an instance of conditional indelpendence) did not appear successful. From this point of view the data are especially suited to modelling involving a hidden variable. Among variety of models with a hidden variable of particular importance would be functional nonparametric models that allow data-based identification of reaction probabilities as functions of the hidden variable, as well as prediction of the values of hidden variable (Kutylowski, 1994).

Figure 5 about here: see consequtive pages with graphics following the end of the draft.

Figure 5: Full model of attributes defining a "true" Japanese: an inventory highly suitable for simplification via a model with a hidden variable.

Essential results from application of such a model to comparative identity of Germans and Poles in 1998 will be provided.

5 Conclusions

Simple tools have been discussed for the purpose of visualizing cross-country comparisons for single polytomous response: bar charts based on category probabilities, cumulative probabilities and sequential probabilities. Special type of quantities, continuation ratios, arise within the sequential approach under the assumption that one uses logit link for the purpose of modelling linera structure of the dependency of response on factors.

The cumulative and sequential approaches lead to respective advanced modelling frameworks, which will not be discussed here. In particular, specialization of sequential modelling in terms

[3] Model selection was conducted using package SCD/digram that has been under development by Svend Kreiner, Department of Biostaatistics, University of Copenhagen. Although that package lacks help facility, some documents by the developed distributed with it may be useful in introducing *modus operandi* of that workbench.

of continuation ratios is discussed in Fienberg (1977, 1980), Fienberg and Mason (1978) as well as in Mare (1980). Kutylowski (19??) provides a comparative view with emphasis on sequential approach.

In multiresponse framework three related approaches have been put forward:

1. semi-parametric nonmodelling approach relying on estimation of Davis-Goodman-Kruskal association coefficient γ, either marginally or conditionally, and

2. semi-parametric modelling approach of deriving graphical models as subsets of log-linear models.

3. nonparametric approach with a hidden variable.

First two approaches can be seen as preliminary with respect to functional nonparametric approach (based on kernel smoothing), in the sense of providing information of relevance before “explanation” by recourse to the hidden variable is implemented.

Taken altogether the proposed methods provide incisive and relative simple tools for exploratory cross-country comparisons.

References

Agresti, A. (1984). *Analysis of Ordinal Categorical Data.* New York: Wiley.

Cox, C. (1988). Multinomial regression models based on continuation ratios. *Statist. in Medicine 7*, 435-441.

Davis, J. A. (1967). A partial coefficient for Goodman and Kruskal’s Gamma. *Journal of the American Statistical Association* 69, 174-180.

Fienberg, S. E. (1980). *The Analysis of Cross-Classified Data*. 2nd ed. (1st ed: 1977). MIT Press.

Fienberg, S.E., Mason, W. M. (1978). Identification and estimation of age-period-cohort models in the analysis of discrete archival data. *Sociological Methodology 1979*, 1-67.

Goodman, L. A. (1983). The analysis of dependence in cross-classifications having ordered categories, using log-linear models for frequencies and log-linear models for odds. *Biometrics 39*, 149-60.

Kahn, L. M. Morimune, K. (1979). Unions and employment stability: a sequential logit approach. *Int. Econ. Rev.*, 217-36.

Kutylowski, A. J. (1991). A comparions of models for ordinal response data. Pp. 1-17 in: K. V. Nielsen, ed.; *Symposium i Anvendt Statistik*. Copenhavn: UNI*C.

Kutylowski, A. J. (1992). Sequential models for ordinal response data. Pp. 195-204 in: van der heijden, P. G. M., Jansen, W., Francis, B. and Seeber, G. U. H., eds., *Statistical Modelling*. Amsterdam: Elesvier.

Kutylowski, A. J. (1994). Nonparametric latent factor analysis of occupational inventory data. Pp. 253-266 in: Rost, J. and Langeheine, R., eds. *Applications of Latent Trait and Latent Class Models in the Social Sciences*. New York: Vaxmann.

Lauritzen and Kreiner. (????). Stata program for estimation of partial gamma.

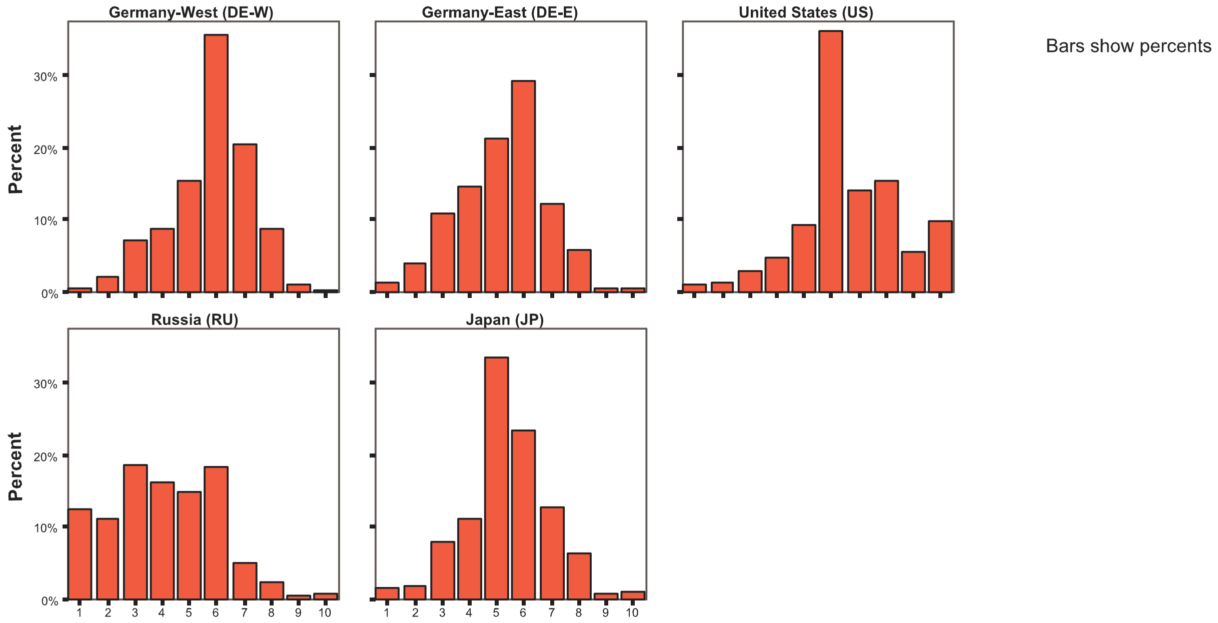
Mare, R. D. (1980). Social background and school continuation decisions. *J. Amer. Statist. Assoc. 75*, 295-305.

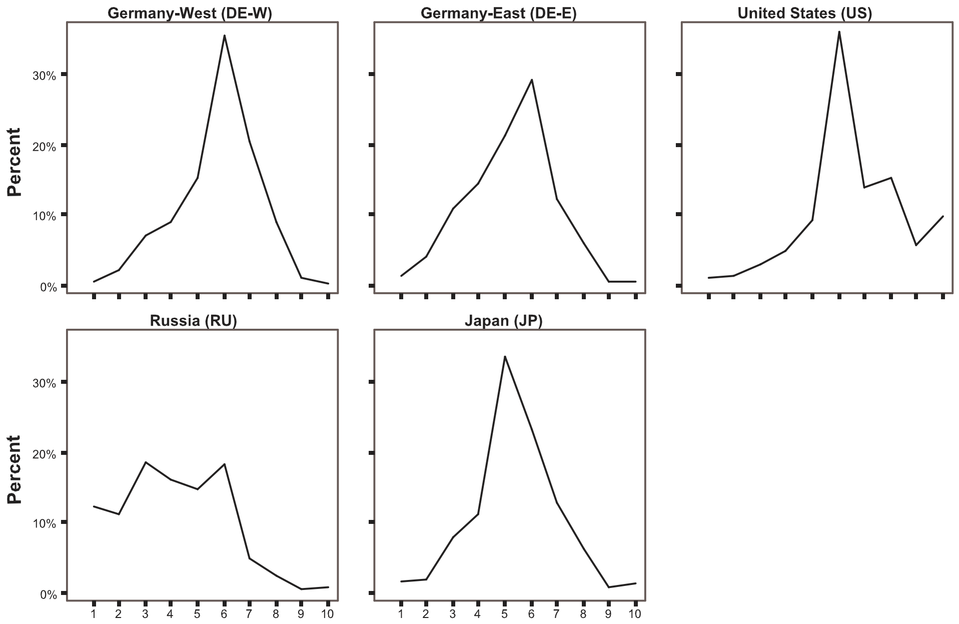
McCullagh, P. (1980). Regression models for ordinal data. *J. R. Statist. Soc.B, 42*, 109-42.

McCullagh, P., Nelder, J. (1989). *Generalized Linear Models*. 2nd ed. London: Chapman and Hall.

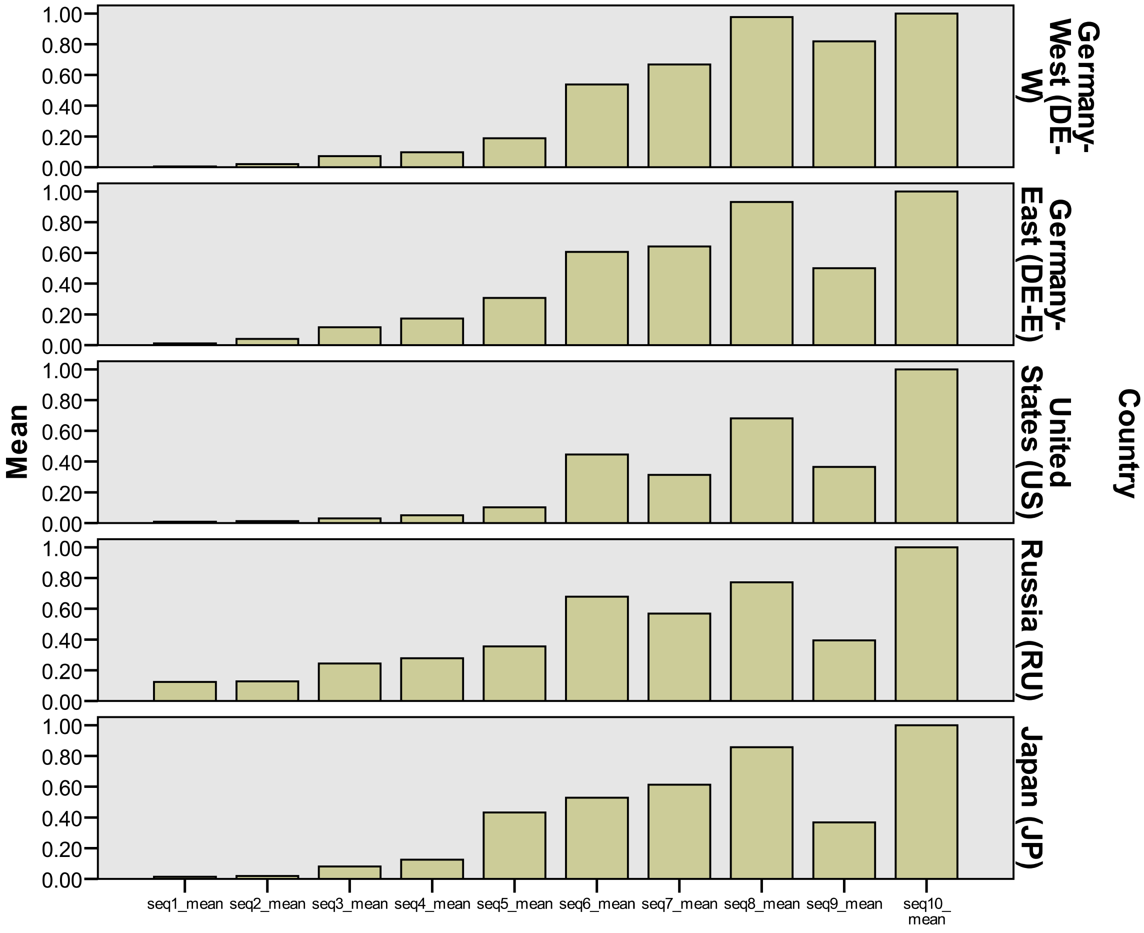
Tutz, G. (1990). Sequential item response models with an ordered response. *Br. J. Statist. and Mathem. Psych.*, in print.

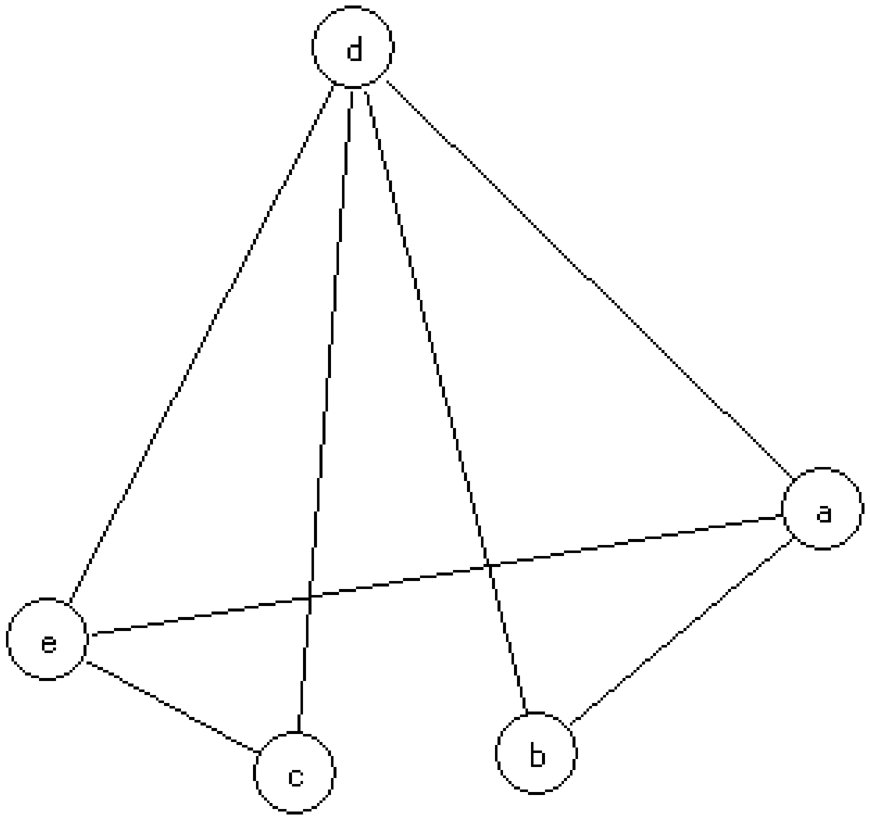
Tutz, G. (1991). Sequential models in categorical regression *Comp. Statistics and Data Analysis 11*, in print.

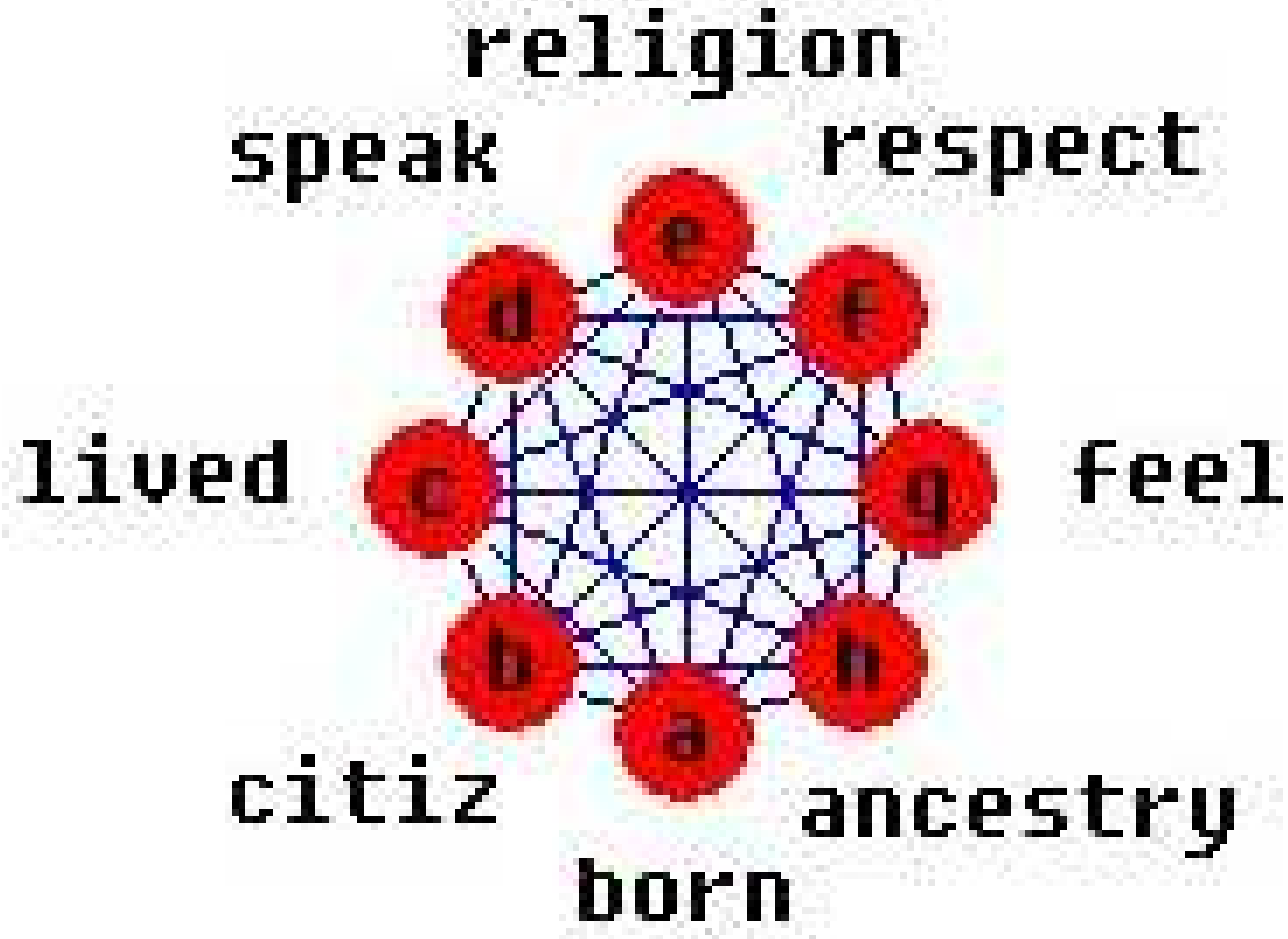

Germany-West (DE-W)
Germany-East (DE-E)
United States (US)
Russia (RU)
Japan (JP)
Percent
Percent
30%
20%
10%
0%
1
2
3
4
5
6
7
8
9
10
Bars show percents



Dot/Lines show percents


Mean
Country
Germany-West (DE-W)
Germany-East (DE-E)
United States (US)
Russia (RU)
Japan (JP)
0.00
0.20
0.40
0.60
0.80
1.00
seq1_mean
seq2_mean
seq3_mean
seq4_mean
seq5_mean
seq6_mean
seq7_mean
seq8_mean
seq9_mean
seq10_mean


d
a
e
b
c


religion
speak
respect
e
d
f
lived
c
g
feel
b
h
a
citiz
ancestry
born