

Standards setting when standardizing categorical data

Paul S. Lambert¹, Jesse M. Blum², Alison Bowes¹, Vernon Gayle¹, Simon B. Jones², Richard O. Sinnott³, Koon Leai Larry Tan², Kenneth J. Turner², Guy C. Warner²,

¹Department of Applied Social Science, University of Stirling, UK; ²Department of Computing Science and Mathematics, University of Stirling, UK; ³National e-Science Centre, University of Glasgow, UK

Email address of corresponding author: paul.lambert@stirling.ac.uk

Abstract. Quantitative social survey research struggles to reconcile the widespread collection of categorical measures, and the common desire to conduct analyses which are not well suited to categorical data (involving comparing standardized, relative positions). By describing approaches taken in the DAMES NCeSS research Node, which is developing facilities to assist in the management of categorical data on occupations, educational qualifications, and ethnicity, this paper argues that tools and practices associated with e-Social Science offer an opportunity to raise standards in the analysis of categorical data.

Introduction

This paper concerns the problem of making meaningful comparative statements concerning social science research data which is collected over more than one context (such as different time periods or countries) but which is recorded in terms of qualitative categories which are not, overtly, equivalent. Does a qualification of ‘higher degree’ obtained in 2008 have the same social significance as one obtained in 1958, for instance? Or how could the UK’s 2001 Census ethnic group category of ‘Black Caribbean’ possibly be compared with the 1972 General Household Survey’s category of ‘Coloured’ (cf. Platt et al 2005)?

Quantitative data collected for social research (such as responses collected from questionnaire surveys) is conventionally regarded as either ‘metric’ or ‘categorical’ (see Stevens, 1946). Categorical measures contain information that concerns a concept which is measured in a manner which is defined according to distinctive boundaries between groups (categories). For a categorical variable, the measurement tells us which mutually exclusive category a response is located in. In practice, far more social research data is categorical in character than metric. Numerous statistical techniques for analysing categorical data are available (e.g. Agresti, 2002). Binary categorical measures (which only distinguish two groups) are well suited to a host of convenient statistical comparisons (e.g. Morgan and Winship 2007). However, multi-category measures (involving three or more categories) are harder to incorporate in statistical analyses. Applications often convert multi-category data into a series of binary contrasts (known as dummy coding; see for example Hardy and Reynolds, 2004); or, less commonly, they approximate a metric measure from the categorical groups (often in substantively implausible ways – see a critical discussion in Angrist and Krueger, 1999).

Managing categorical data

Decisions concerning the allocation and ‘recoding’ of categorical data are a major part of the social science workflow (Long, 2009). Typically, expansive multi-category data is collected, but for analytical purposes this is recoded into a much smaller range of categories. Common recoding processes are unsatisfactory for three reasons. Firstly, decisions are often made in an *ad hoc* manner with limited consistency with previous research and which fail to exploit suitable information resources. We have written extensively about problems of this nature associated with occupational data (e.g. Lambert et al. 2007), and data on educational qualifications and ethnicity (Lambert et al., 2008). Such practices may well reflect lack of expertise in the component topic, which naturally occurs in multivariate research (as a recent example, compare Khattab’s 2009 thoughtful coding of ethnicity with the simpler coding of occupations from the same analysis). Secondly, the analytical benefits of recoding categories are questionable, since techniques do exist to give parsimonious results for complex categorical data (feasible, under-used analytical approaches include regarding categories as clusters which can be modeled as random effects - see examples for occupations in McGovern et al. 2007; or fitting parameters for the direct effects of large numbers of categories using devices such as Stata’s ‘areg’ – see StataCorp, 2009). Third, decisions are seldom well documented and replicable by other researchers. Documentation and replicability is highly desirable, but despite increasing internet- and software-based opportunities for the documentation of social survey analysis (Dale, 2006; Long, 2009), little social science research is reported in a replicable manner.

Such activities are examples of ‘data management’ applied to categorical data. The DAMES Node (www.dames.org.uk) is developing services to improve standards in data management generally, and, in this example, in dealing with categorical data. Key contributions involve providing facilities for documenting operations such as recoding commands; for characterizing and coordinating sequences of activities through workflows; and by providing guidance on accessing and exploiting the heterogeneous external data resources which may be relevant to categorical data. In a recent technical paper, two of the current authors used such methods to show alternative means of comparisons and standardizations for categorical data associated with the UK’s Research Assessment Exercise, an evaluation of higher education research quality (see Lambert and Gayle, 2008, and RAE, 2008). In that paper, we discussed how categorical data (on Higher Educational Institutes, Units of Assessment, and Grade Point Rankings) could be processed and analysed in a manner which facilitated comparisons over time, between institutions, and between subject areas. Moreover, we argued that previous analyses of the RAE results had employed unjustifiably simplified analytical approaches, given that more sophisticated comparisons could readily be made (with appropriate facility in ‘data management’ regarding the preparation and manipulation of complex data resources).

Standardizing categorical data

Two approaches are commonly employed to analysing categorical data for the purposes of comparative research (making comparisons across different contexts such as different countries or time periods). Both are referred to as ‘standardizing’ or ‘harmonizing’ data, but involve very different processes and assumptions about equivalence (Ehling, 2003).

Measurement equivalence

The idea of ‘measurement equivalence’ is to control the collection and/or analysis of categorical data in order that, for analytical purposes, the information captured by a given

categorisation is directly comparable across contexts (van Deth, 2003). Measurement equivalence allows for the most convenient form of (direct) comparison. Control or 'harmonisation' during the data collection process is (in principle) achieved by instruction from survey designers, national statistics agencies, and other forms of academic quality monitoring (Ehling, 2003; Harkness, 1999).

Yet in practice, very few social science measures can be uncritically agreed as achieving measurement equivalence across contexts. The questions posed at the start of this text, concerning the consistency of educational and ethnic group categories over time, are good examples where measurement equivalence is either not possible (because the data was collected in different categories, such as on ethnicity), or not plausible (because in most measures, the circumstances of a PhD qualification in 1959 and 2009 would not be considered equivalent). The common critique is that calling a category the same entity across contexts may achieve a token harmonisation in name, but little else (Ehling, 2003). In our recent analysis of the RAE 2008 results, for instance, we show that it is empirically implausible to accept the assertion of meaning equivalence which underpins most reporting on the RAE (namely that the categorical ratings applied by different RAE subject panels are consistent across disciplines, see Lambert and Gayle, 2008). Further extended critical discussions of attempts at establishing measurement equivalence can be found in specialist literatures such as cross-national comparisons regarding socio-demographic variables (e.g. Hoffmeyer-Zloknik and Wolf, 2003).

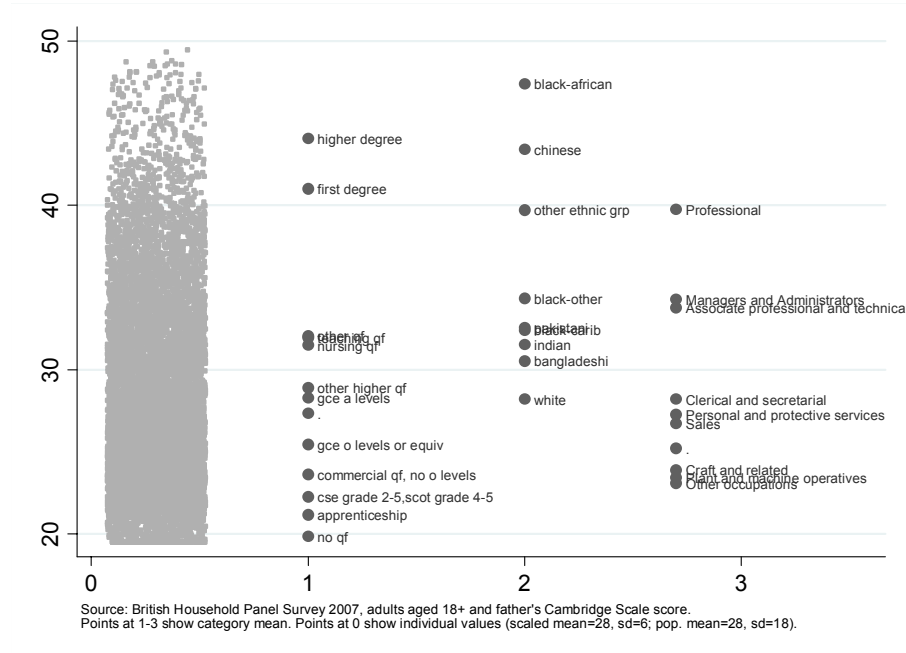
Measurement equivalence might, however, be achieved by detailed processing and recoding of data after data collection ('*ex post*'). This task is usually aided by expanding the number of category options within the data collection taxonomies, and putting considerable effort into assigning detailed records into suitable analytical categories. This is the most common strategy in contemporary survey research, and is the focus of most existing resources directed towards the measurement and analysis of categorical variables in the social sciences (e.g. ESDS, 2008; ONS, 2005). There are numerous promising initiatives in achieving *ex post* measurement equivalence, such as extended reviews of particular measurement instruments which achieve effective validation of equivalence for data on educational qualifications (Schneider, 2008; Brauns et al., 2003) and on occupations (Ganzeboom and Treiman, 1996). Progress in these aims is widely regarded as contingent upon communication of appropriate data collection and recoding strategies, and agreement upon appropriate terminologies and underlying concepts. Nevertheless, critics of the measurement equivalence approach highlight the unpersuasive tendency of '*ex post*' equivalence efforts tend to harmonise to the lowest common denominator in terms of detailed difference between categories, apparently in order to keep the number of cases within categories high and the range of categories simple; the loss of local detail which can be associated with harmonised measures (e.g. Lemel, 2002); and the artefactual disjunctures between the distributions of measures in different contexts which can arise from institutional variations or revisions in recording and recoding categories.

Meaning equivalence

The notion of meaning equivalence invokes the comparison of locations according to their relative, contextual meaning (e.g. van Deth, 2003). As an illustration, the relative social advantage associated with having a PhD level qualification may be substantially less in the United Kingdom than in Ghana (say), so an analysis prioritising meaning equivalence would regard PhD's held in the two countries as marking *different* relative levels of educational attainment. For categorical data, such contextualised measures are not commonly employed, yet the relative positions of categories can be readily be explored by undertaking a statistical analysis of the categories' position(s) within dimension(s) of difference.

Dimension scores can be obtained from simple summary statistics from related data (such as averages across groups), or techniques such as ‘correspondence analysis’ or ‘stereotyped ordered logistic regression models’ which derive scores in a multivariate framework (for examples see Lambert, 2005). Figure 1 shows an illustrative representation of assigning dimension scores to a modestly complex number of categories in three domains. In Figure 1, a recognisable ranking of occupational categories and educational qualifications is shown. The scores associated with these ranks could themselves be used as measures of the relative advantage typically associated with these categories. The ranking of ethnic group categories in Figure 1 is less recognisable. The ranking seems to suggest that all minority groups have, on average, relatively more advantaged parental occupations (which is plausible given immigration criteria and the skewed occupational distribution of immigrants). In the absence of additional analyses, the scores of those groups might be interpreted as markers of relative advantage conditional on immigrant background.

Figure 1: Dimension scores: Ranking categories by ‘Father’s occupational advantage’



Dimension scores such as illustrated in Figure 1 offer a measure of the categories’ position within a wider structure. In social science research there are traditions of assigning dimension scores to occupational unit groups in order to derive ‘prestige’ or ‘stratification’ scales (e.g. Treiman, 1977; Rytina, 1992; www.camsis.stir.ac.uk). However there are few other initiatives in scaling categorical data for these purposes (it should be noted that scaling in more than one dimension is potentially useful, as is the analysis of dimensions in non-linear functional forms, though in practice neither are commonly found in social science applications). Scaling is attractive both for its intrinsic research interest, and for the convenient analytical properties of scales. The latter are most important because, first, dimensionalising categories largely removes the pressure to merge categories due to sparse representation, and can thus conveniently sustain much larger numbers of different categories; second, because dimensionalising categories allows for easier operationalisation of main and interaction effects in multivariate analyses (since a linear or curvilinear effect is modelled, rather than a series of dummy variables); and lastly, because dimensional measures are readily adapted to arithmetic standardisation. An arithmetic standardization refers to techniques used for comparison of statistical summaries of different measures on a coordinated scale. A common standardization for metric data involves re-scaling the data around the mean and standard

deviation of the distribution. A value on a standardized variable therefore tells us where the response fits in terms which are relative to the distribution of the wider population. This type of standardization is extremely helpful in making comparisons between different populations in which the original scales do not have identical distributions. Arithmetic standardization thus offers a convenient form of ‘meaning equivalence’.

Supporting standardization of categorical data

A challenge for an analysis interested in either measurement or meaning equivalence is the plethora of alternative treatments for categorical data which might be suitable. Numerous recoding tools and instructions are available, and numerous potential dimension scores may already exist or might be calculated.

Existing resources for categorical data in the social sciences

A great many resources exist which can be effectively used by social survey researchers to enhance their use of categorical data for the purposes of comparative research. Several altruistic academic resources can be identified, such as personally maintained websites and information resources covering suggested coding and analytical approaches, themed according to specialist topics or data resources (e.g. Ganzeboom, 2008). There are also major national and international initiatives in supporting information on measures and their analysis in survey research. Selective examples include ESDS (www.esds.ac.uk), supporting analysis of UK oriented datasets which includes guidance papers and websites on generic topics in data management (Raferty and Watham, 2008), and on specialist information on the nature of selected categorical measures, and advice on their harmonization over time and between countries such as in software scripts (ESDS, 2008); the UK ONS Harmonisation Unit, which publishes online and text resources on measuring and categorizing key contemporary measures (ONS, 2008); and resources associated with particular data collections such as IPUMS (www.ipums.org), LIS (www.lisproject.org) and the European Social Survey (www.europeansocialsurvey.org). The scale of these resources far exceeds those that can be developed by the DAMES Node, yet we argue nevertheless that there is scope for an additional contribution from that Node. This arises because the standard model of information provision associated with existing resources is static and hierarchical in which expert advice is given over recoding approaches. We observe that much research proceeds without adhering to such advice. This is apparently because users regard the instructions either as too difficult to follow (recommending operations that the user is not confident to implement); unsatisfactory (recommending a crude or unstandardised measure); out of date (referring to a time period which is too late or too early for the data); or because users are simply not aware of the relevant data resource.

DAMES resources for recoding data and dimension scores

We argue that an e-Science approach can begin to address the lack of impact of existing resources, by seeking to offer seamless access to plural, dynamic heterogeneous resources relevant to standardising categorical data. The DAMES Node (www.dames.org.uk) is seeking to achieve this objective on two fronts. The first involves developing capacity for complex data management operations associated with standardizing categorical data. This partly involves software oriented support, where an effective package such as Stata supports rich data manipulations on complex categorical data and well-documented command logs (Long, 2009). More general contributions concern tools supporting documentation and replication of data management tasks, workflow characterisations of such tasks, and communication of related complex activities. The DAMES Node is contributing here by pursuing tools,

compatible with major data resource providers, which record metadata in standard formats (DDI 3, see Vardigan et al., 2008) concerning key tasks in managing categorical data (concerned with 'operationalising variables' and 'linking data files'). Some of these tools are oriented to specialist data resources (on occupations, education, ethnicity); others are being designed in a generic format to facilitate user-friendly data manipulation and documentation.

The Node's second contribution involves enhancing data resources concerned with categorical data. It seeks to improve the fluency with which categorical data is linked to other information about those categories for the purposes of analysis. That information may cover recommended recoding syntax (as advocated by a major statistics institute) but it might equally involve more contestable information which is supplied from other sources (such as academic research projects). Such information is voluminous and heterogeneous. The GE*DE project within the DAMES research Node is aiming to promote such resources by allowing, via online Portals, easy access to, and depositing of, specialist information resources concerned with categories of occupational records, educational qualifications, and ethnicity / immigration-related social group (see Lambert et al., 2008). At time of writing, the GE*DE services provide access to numerous resources for occupations, and smaller numbers for educational qualifications and ethnicity. Over time, this resource should allow easier access to these data to large numbers of users. The DAMES project also includes training and capacity building activities designed to encourage social scientists to exploit this data.

A significant contribution to the standardisation of categorical data involves both generating, and disseminating, dimension scores for categories, in order to support scaling according to relative position and the concordant approach of meaning equivalence. Such scaling is well catered for in the field of research on occupations, but is much less widely employed for data on educational qualifications and ethnicity. We argue that these domains for social science research are ideally suited to dimension scaling since they are pervaded by important categorical divisions which are lost in categorical harmonisations associated with measurement equivalence (in research associated with educational qualifications, this primarily concerns treating qualifications within different birth cohorts as qualitatively different; in research on ethnicity and migration, this primarily concerns cross-classifying different socially significant ethnic referents, such as data on identity, national origins, religion, language - cf. Khattab, 2009).

Conclusions: Contributions towards good practice

Given the descriptions above, it can be claimed that there are two contributions which can be made to improving the management of categorical data for the benefit of comparative research: drawing researchers towards suitable supplementary data on their categorical measures which they were not previously aware of; and drawing them towards best practice analyses and manipulations of that data which they were not previously aware of or confident with. In particular, we argue that for the purposes of statistical analysis, ranking according to relative position within a dimension of difference is a popular and effective means of standardisation for comparative research, and an important contribution could be made to ensure suitable scaling measures are easily accessible and widely understood.

An e-Science orientation can help address the shortcomings of existing analyses of complex categorical data by supporting the systematic collection and storage of suitable metadata on categorical data; the development of mechanisms allowing access to specialist external information resources which may be relevant to particular measures; and the facilitation of complex analytical techniques suited to categorical measures. All of these approaches offer contributions to the analysis of complex categorical data which are effectively approached through the framework of e-Social Science.

Acknowledgments

The NCeSS DAMES Node is supported by the UK ESRC, grant reference RES-149-25-1066.

References

- Agresti, A. (2002). *Categorical Data Analysis, 2nd Edition*. New York: Wiley.
- Angrist, J. D., & Krueger, A. B. (1999). Empirical strategies in Labour Economics. In O. C. Ashenfelter & D. Card (Eds.), *Handbook of Labour Economics, Vol. 3* (pp. 1277-1366). Amsterdam: Elsevier.
- Brauns, H., Scherer, S., & Steinmann, S. (2003). The CASMIN Educational Classification in International Comparative Research. In J. H. P. Hoffmeyer-Zlotnik & C. Wolf (Eds.), *Advances in Cross-National Comparison: A European Working Book for Demographic and Socio-Economic Variables* (pp. 221-244). New York: Kluwer Academic.
- Dale, A. (2006). Quality Issues with Survey Research. *International Journal of Social Research Methodology*, 9(2), 143-158.
- Ehling, M. (2003). Harmonising Data in Official Statistics: Development, Procedures, and Data Quality. In J. H. P. Hoffmeyer-Zlotnik & C. Wolf (Eds.), *Advances in Cross-National Comparison* (pp. 17-31). New York: Kluwer.
- ESDS. (2008). Occupational coding in the LFS. Retrieved 1 May, 2009, from <http://www.esds.ac.uk/government/dv/nssec/lfs/>
- Ganzeboom, H. B. G., & Treiman, D. J. (1996). Internationally Comparable Measures of Occupational Status for the 1988 International Standard Classification of Occupations. *Social Science Research*, 25(3), 201-235.
- Ganzeboom, H. B. G. (2008). Tools for deriving status measures from ISKO-88 and ISCO-68. Retrieved 1 March, 2008, from <http://home.fsw.vu.nl/~ganzeboom/PISA/>
- Hardy, M., & Reynolds, J. (2004). Incorporating Categorical Information into Regression Models: The Utility of Dummy Variables. In M. Hardy & A. Bryman (Eds.), *Handbook of Data Analysis* (pp. 210-236). London: Sage.
- Harkness, J. (1999). In pursuit of quality: issues for cross-national survey research. *International Journal of Social Research Methodology*, 2(2), 125-140.
- Hoffmeyer-Zlotnik, J. H. P., & Wolf, C. (Eds.). (2003). *Advances in Cross-national Comparison: A European Working Book for Demographic and Socio-economic Variables*. Berlin: Kluwer Academic / Plenum Publishers.
- Khattab, N. (2009). Ethno-religious Background as a Determinant of Educational and Occupational Attainment in Britain. *Sociology*, 43(2), 304-322.
- Lambert, P. S. (2005). Ethnicity and the Comparative Analysis of Contemporary Survey Data. In J. H. P. Hoffmeyer-Zlotnik & J. Harkness (Eds.), *Methodological Aspects in Cross-National Research* (pp. 259-277). Mannheim: ZUMA-Nachrichten Spezial 11.
- Lambert, P. S., & Gayle, V. (2008). *Data management and standardisation: A methodological comment on using results from the UK Research Assessment Exercise 2008*. Stirling, University of Stirling: Technical Paper 2008-3 of the Data Management through e-Social Science Research Node (www.dames.org.uk).
- Lambert, P. S., Gayle, V., Tan, K. L. L., Blum, J. M., Bowes, A., Jones, S., Turner, K. J., Warner, G., Sinnott, R. O., & Bihagen, E. (2008). *Grid Enabled Specialist Data Environments: Forward Planning for the GE*DE Services for Specialist Data on Occupations, Educational Qualifications, and Ethnicity*. Stirling, University of Stirling: Technical Paper 2008-1 of the Data Management through e-Social Science research Node (www.dames.org.uk).
- Lambert, P. S., Tan, K. L. L., Turner, K. J., Gayle, V., Prandy, K., & Sinnott, R. O. (2007). Data Curation Standards and Social Science Occupational Information Resources. *International Journal of Digital Curation*, 2(1), 73-91.
- Lemel, Y. (2002). Social Stratification: The Distinctiveness of French Research. In Y. Lemel & H. H. Noll (Eds.), *Changing Structures of Inequality: A Comparative Perspective* (pp. 17-44). Montreal: McGill-Queens University Press.
- Long, J. S. (2009). *The Workflow of Data Analysis Using Stata*. Boca Raton: CRC Press.

- Morgan, S. L., & Winship, C. (2007). *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge: Cambridge University Press.
- ONS. (2007). National Statistics Harmonisation. Retrieved 1 May, 2009, from <http://www.statistics.gov.uk/about/data/harmonisation/>
- Platt, L., Simpson, L., & Akinwale, B. (2005). Stability and change in ethnic groups in England and Wales. *Population Trends*, 121, 35-46.
- RAE. (2008). RAE 2008: Research Assessment Exercise. Retrieved 18 December, from <http://www.rae.ac.uk/>
- Rafferty, A., & Wathan, J. (2008). *Working with survey files: using hierarchical data, matching files and pooling data*. Manchester: Economic and Social Data Service, and <http://www.esds.ac.uk/government/resources/analysis/>.
- Rytina, S. (1992). Scaling the Intergenerational Continuity of Occupation : Is Occupational Inheritance Ascriptive after all? *American Journal of Sociology*, 97(6), 1658-1688.
- Schneider, S. L. (2008). *The International Standard Classification of Education (ISCED-97). An Evaluation of Content and Criterion Validity for 15 European Countries*. Mannheim: MZES.
- StataCorp. (2009). Stata Statistical Software, Release 10.1. College Station, TX: StataCorp LP.
- Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, 103, 677-680.
- Treiman, D. J. (1977). *Occupational Prestige in Comparative Perspective*. New York: Academic Press.
- Treiman, D. J. (2009). *Quantitative Data Analysis: Doing Social Research to Test Ideas*. New York: Jossey Bass.
- van Deth, J. W. (2003). Using Published Survey Data. In J. A. Harkness, F. J. R. van de Vijver & P. P. Mohler (Eds.), *Cross-Cultural Survey Methods* (pp. 329-346). New York: Wiley.
- Vardigan, M., Heus, P., & Thomas, W. (2008). Data Documentation Initiative: Towards a Standard for the Social Sciences. *International Journal of Digital Curation*, 3(1), 107-113.