

Collaborative systems for enhancing the analysis of social surveys: the Grid Enabled Specialist Data Environments

Paul Lambert¹, Guy Warner¹, Tom Doherty², Susan McCafferty², John Watt², Michael Comerford², Vernon Gayle¹, Larry Tan¹, Jesse Blum¹, Alison Bowes¹

¹University of Stirling, UK, contact e-mail: paul.lambert@stirling.ac.uk

²University of Glasgow, UK

Abstract

This paper describes a group of online services which are designed to support social survey research and the production of statistical results. The 'Grid Enabled Specialist Data Environment' (GESDE) services constitute three related systems which offer facilities to search for, extract and exploit supplementary data and metadata concerned with the measurement and operationalisation of survey variables. The services also offer users the opportunity to deposit and distribute their own supplementary data resources for the benefit of dissemination and replication of the details of their own analysis.

The GESDE services focus upon three application areas: specialist data relating to the measurement of occupations; educational qualifications; and ethnicity (including nationality, language, religion, national identity). They identify information resources related to the operationalisation of variables which seek to measure each of these concepts - examples include coding frames, crosswalk and translation files, and standardisation and harmonisation recommendations. These resources constitute important supplementary data which can be usefully exploited in the analysis of survey data. The GESDE services work by collecting together as much of this supplementary data as possible, and making it searchable and retrievable to others. This paper discusses the current features of the GESDE services (which have been designed as part of a wider programme of 'e-Science' research in the UK), and considers ongoing challenges in providing effective support for variable-oriented statistical analysis in the social sciences.

Keywords: Survey variables; metadata; e-Science

1. Introduction

1.1. Survey variables and their statistical analysis

All statistical results are ultimately identified from empirical patterns involving manifest variables, and in social survey research it is well recognised that the qualities of an analysis hinge critically upon the means through which variables are 'operationalised' or made manifest (e.g. Bulmer, 1956; Cronbach et al., 1972). Indeed, there is considerable

negotiation over the manifestation of measures, and it is ordinarily the case that several different variable operationalisations are plausible. Popular scientific models for social research promote the comparative evaluation of multiple operationalisations, and the cumulative exploitation of previously used standards by replication based upon the documentation of earlier studies (cf. Dale, 2006; Freese, 2007). Yet social science research traditions lack agreement and consistency over optimal operationalisations, and analysts commonly lack the tools or resources to replicate a variety of relevant measures.

The analysis of data on educational qualifications provides an example. Numerous different means of operationalising measures of educational qualifications are available, ranging across various different categorical classifications, and scaling approaches which assign scores to categories (e.g. Buis, 2010; Schneider, 2008). Table 1 shows results on the correlation between six popular measures of educational qualifications in UK research, and three other important variables. We see substantial patterns of variation between measures: it should not be presumed that analysis using each of the six measures will always lead to the same conclusions.

<i>Education measures</i>	(1)	(2)	(3)	(4)	(5)	(6)
(1) UK scheme in 4 categories		1.00	1.00	0.47	0.62	0.31
(2) Binary (degree)	0.32		0.07	0.21	0.28	0.19
(3) Binary (low or none)	0.37	0.09		0.20	0.32	0.18
(4) ISCED (8 category)	0.72	1.00	0.83		1.00	0.36
(5) ISCED (3 category)	0.48	0.68	0.67	0.51		0.30
(6) School leaving age	0.14	0.24	0.21	0.11	0.17	
<i>Other relevant measures</i>						
Age (linear)	0.045	0.018	0.112	0.069	0.061	0.118
Gender	0.003	0.001	0.004	0.003	0.004	0.000
Father's occ. advantage (CAMSIS)	0.049	0.091	0.076	0.037	0.060	0.143
Source: Analysis of the UK's British Household Panel Survey (Univ. Essex, 2010), adults aged 16+ in Wave R (2008). Pairwise N approx. 11500-12500; analysis uses cross-sectional weights. Values are bivariate (pseudo)r ² from multinomial logit or linear regression. Correlations asymmetric due to predictive logic. See www.dames.org.uk/geede for information on derivations of measures of educational attainment.						

In an extended review, Schneider (2008, 2010) documents numerous difficulties in recording and analysing data on educational qualifications for cross-national comparative research – for instance, institutional differences between societies cannot easily be reconciled into comparable educational categories, and applied research exhibits many different strategies involving coding measured categories (see also Hoffmeyer-Zlotnik and Warner, 2005). Schneider notes that some efforts have been made to standardise the measurement of educational data, but these are overwhelmingly rejected, or ignored, by applied researchers, often for sound academic reasons of comparison (cf. Chauvel, 2002). In the face of resistance to standardisation, a ‘bottom-up’ strategy, providing researchers with easy access to a full range of suitable operationalisation possibilities, might seem the most plausible scientific approach; our own table 1 shows the potential benefits of sensitivity analysis in this style. Yet in practical terms very few researchers seem readily able to adopt and compare several different alternative measures. In an information-rich age, the collaborative exchange of information on social survey variable operationalisations seems surprisingly ineffective.

1.2. e-Science and the GESDE services

Over the last decade a series of applications of 'e-Science' approaches to social science research have been developed, fostered by initiatives such as the UK's programmes in 'e-Social Science' (Halfpenny et al., 2009) and 'Digital Social Research' (e.g. www.digitalsocialresearch.net/). Broadly, these applications promote collaboration and communication for research purposes; the improved sharing and organisation of data resources; and facilities for analysing and summarising complex data and its results.

This paper describes one group of e-Science services which are designed to improve standards in social survey research. The 'Grid Enabled Specialist Data Environment' (GESDE, see www.dames.org.uk/themes.html) online services constitute three related 'portal' systems which offer users facilities to search for, extract and exploit supplementary data and metadata concerned with the measurement and operationalisation of survey variables (see further descriptions on the website). The services also offer users the opportunity to deposit and distribute their own supplementary data resources for the benefit of dissemination and replication of the details of their own analysis.

The GESDE services focus upon three particular application areas: specialist data relating to the measurement of occupations ('GEODE'); educational qualifications ('GEDE'); and ethnicity ('GEMDE', covering data on nationality, language, religion, and identity). Information resources related to the operationalisation of variables which seek to measure these concepts are stored - examples include coding frames (e.g. Ganzeboom, 2011), crosswalk and translation files (e.g. Leiulsfrud et al., 2005), and standardisation and harmonisation recommendations (e.g. ONS, 2010). These resources constitute important supplementary data which can be usefully exploited in the analysis of survey data. Nevertheless, in published research, many survey researchers do not take full advantage of potentially relevant supplementary data, perhaps because they are simply not aware of its existence. The GESDE services seek to collect together as much of this supplementary data as possible, and making it searchable and retrievable to other users.

The GESDE services are motivated by recognition of the centrality of 'data management' activities to the process of undertaking survey analysis and generating statistical results. Firstly, the tasks of data management themselves occupy a major component of most research projects, but the coordination and collaboration possible using the GESDE services might reduce such demands substantially. Secondly, many different data management choices can be taken, involving different codings and standardisations of measures, and these can lead to different statistical results (as Table 1); accordingly, the GESDE services can be used to support sensitivity analysis involving comparisons across different measurement options, including the replication of those used in previous studies. This is hoped in the long term to result in analytical results which are more robust, better documented, and more cumulative in their relation to previous research.

The GESDE services originate from a UK project on 'Data Management through e-Social Science' (www.dames.org.uk) which has a general focus on tasks involving adjusting and enhancing social science research data. They are also implemented within two other UK-

based projects with a focus on the production of social statistical results (the e-Stat research Node, <http://www.cmm.bristol.ac.uk/research/NCESS-EStat/>, at the University of Bristol, which is designing a facility for building statistical models of complex social processes on complex data structures, designed to be tractable and replicable to non-specialist users; and the National e-Infrastructure for Social Simulation, NeISS, <http://www.neiss.org.uk>, see Birkin et al. 2010, led at the University of Leeds, which is building a facility to prepare data, specify models, and summarise results from a number of different social simulation modeling routines). The GESDE services are currently supported via servers at the Universities of Glasgow and Stirling, though in the long term these could be relocated or maintenance arrangements otherwise adjusted.

2. Implementation of the GESDE services

The GESDE services are made available through online ‘portal’ environments. These are secure environments which provide access to an integrated set of services. They support facilities to:

- login with different levels of access permission, covering a ‘guest’ level access to detect materials but not to edit or deposit materials, and a ‘named’ user access level where the registered user also has permission to upload, edit and comment on data resources
- search across an existing, dynamic pool of information resources stored in the systems, and navigate search results according to criteria concerning the resources (such as date or user rating)
- enter new data into the system, by submitting structured metadata through a standardised form and by uploading one or more information files
- submit rating information on the quality of particular resources
- download data from the system through a variety of user-friendly means
- allow expert user groups (such as members of the project) to monitor, and to edit, update or remove submitted records (such as for quality enhancement purposes)
- perform analytical operations using the GESDE data resources, such as to generate statistical summary data from stored files, or to merge together the user’s local data with a suitable remote ‘donor’ dataset.

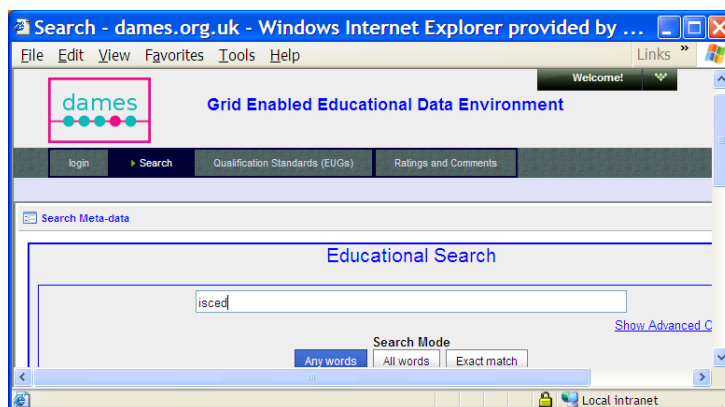


Figure 1a: The ‘search’ facility for a guest level user on ‘GEEDE’



Figure 1b: Results available on GEDE after searching for 'ISCED'

Figures 1a, 1b, 2 and 3 illustrate a selection of these functionalities. Figure 1a shows a guest level user having accessed the GEDE service ('Grid Enabled Educational Data Environment') and preparing to run a search on the term 'ISCED'. After running the search, they are presented with a list of matching data resources stored in the system. The full record, with links to downloadable resources, for one result is shown in Figure 1b.

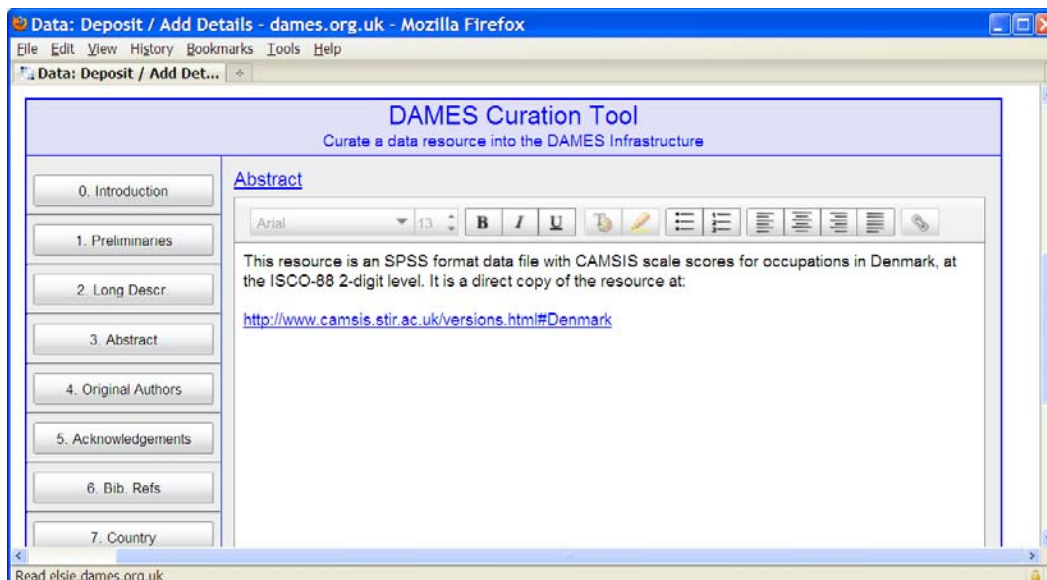


Figure 2: 'Curating' an occupational information resource at GEODE

Figure 2 shows a depiction of the ‘data curation tool’ used within the GEODE and GEEDE services to collect systematic metadata on the information resource being uploaded – the image shows the author adding an abstract to describe their data resource. Lastly, figure 3 shows an image from the ‘microdata analysis’ service within the GEMDE service (‘Grid Enabled ethnic Minority Data Environment’). This service involves running a bespoke query on a survey micro-data set held on the server. This functionality is very similar to that provided by tools such as NESSTAR (cf. Rafferty and Smith, 2008), but the contribution to the GESDE services is that highly focused results are available which are pertinent to understanding the specialist topic involved (in this example, a regression model for ‘ethnic penalties’ is estimated according to the requested parameters, following the analytical standard set by Heath and Cheung, 2007).

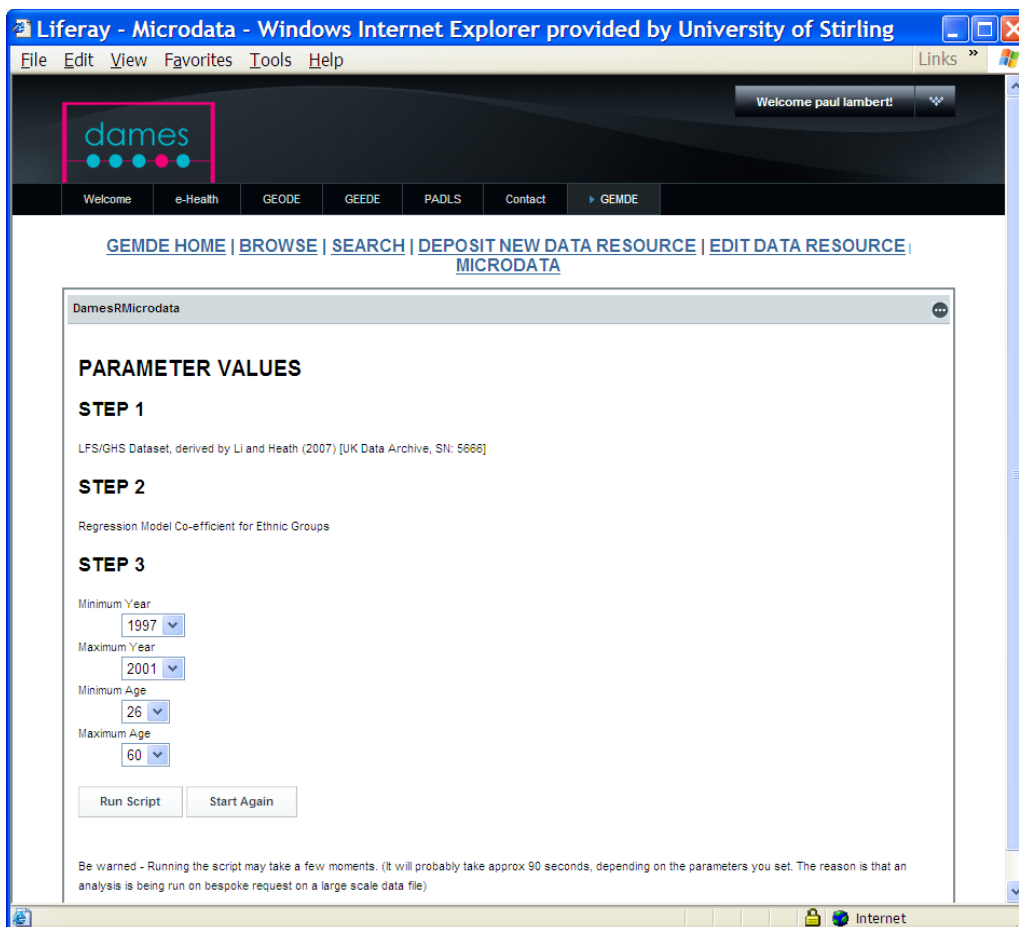


Figure 3: Specification of job using the GEMDE ‘microdata analysis’ tool

More extended guides to using the GESDE data resources are published on the webpages for the sites (see www.dames.org.uk/themes). As applications of e-Science services, there are several noteworthy components to the GESDE services. Each portal connects dynamically to data storage facilities which host data submitted to the systems - iRODS software is used for this purpose (www.irods.org). The organisation of metadata submitted by the depositors of resources is critical to the storage arrangements – DDI format metadata is stored to support easy extraction of data components in combination

with compatibility with other major social science data resources (Vardigan et al., 2008). Data security is ensured by content management systems which allow delegation of access to named users of data resources as required (by default, new resources entered into the systems are denied to all other users until they have been approved by at least one individual from the project team). Finally, steps are taken to navigating the heterogeneity of resources and formats by providing automated routines for implementing jobs across popular formats and for allowing data depositors (and members of the project team) to add annotations to clarify the character of their data resources.

3. Using GESDE to raise standards in survey statistics

3.1 Problems of design, uptake and quality

The GESDE services ultimately aim to raise standards in survey statistics by enabling breadth and depth in variable operationalisations in research analysis. However substantial practical challenges of design, uptake and quality control arise in successfully providing online services for these purposes. Of the former, whilst work has been ongoing on these services since 2008, there remain various programming bugs or unintended gaps in services. In our own project's experience these arose due to the practical difficulties of developing a non-commercial online service with limited programmer commitments; a requirement to use freeware resources which themselves are regularly updated; and a commitment to supporting a range of services which potentially bring conflicting or inconsistent system requirements. The academic project behind the GESDE service development continues to collect user feedback and seeks to develop solutions as required, but it seems unlikely the all perceived design problems will ever be comprehensively eliminated.

The uptake of the GESDE services is clearly a critical barometer of their contribution, since they are dynamically populated by contributions from researchers themselves. Approaches to incentivise registering data on the GESDE systems have been considered by the project group, centring upon citation rewards for contributing authors, however the process of submitting information to the services, particularly when bugs are found on the developmental interfaces, is demanding. Hitherto, the large majority of resources on the GESDE services have been submitted by members of the GESDE project themselves, and future dissemination and encouragement to people outside this group to submit materials is a clear future priority.

The GESDE services have primarily been designed by, and for, academic social science researchers as part of a project based in the UK. However the systems are open to input from members of other organisations and from other countries. In practice, a large proportion of the data resources in each service are not specific to the UK, and a particularly important collection of resources to the GESDE services are those concerned with harmonising and standardising measures of occupations, educational qualifications and ethnicity for the purposes of comparative research (both between countries and across time): the GESDE services offer one route for researchers to identify and exploit recommended previous approaches to harmonising variables, and to disseminate their own approach if relevant.

With regard to quality control, an attraction of the GESDE services is that they follow a pluralistic, academic approach to measurement options: several different harmonisation recommendations may be available, for instance, including but not restricted to standards recommended by National and International Statistical Agencies. Such a pluralistic approach, however, clearly raises questions over the relative quality standards of different resources - particularly pertinent in the subject domains of the GESDE services there are already literatures addressing quality standards (e.g. Rose and Harrison, 2010). The GESDE services have features which contribute to the evaluation of resources supplied to them, in the form of user-ratings and user-comment tools, and the capacity of nominated 'expert users' to annotate, update, amend or if necessary delete submitted resources. Nevertheless such steps represent a liberal rather than authoritarian approach, which cannot remove all risks related to quality. Appealing again to scientific standards of transparency and documentation, the only truly robust defence of quality concerns which can be offered by the GESDE services is that have an explicit citable identity (linked to the date of curation and the identity of the resource supplier). If the users of resources accessed through GESDE take adequate steps to cite the specific resources concerned, a substantial step is taken towards making the dissemination of resources through these systems more scientifically robust.

3.2 Opportunities for improving statistical analysis of social survey data

A focus on data management in three subject fields (measures of occupational position; educational qualifications; and ethnicity) presents exciting opportunities for improving scientific standards in the analysis of variables. Firstly, as earlier represented through Table 1, it is possible to demonstrate non-negligible differences between the statistical results derived from analysis using what were overtly comparable measures of educational level (for example, the range in values of the correlation statistic with the measure of father's occupational advantage was very substantial – analysis using some measures might suggest a very small and perhaps not statistically significant association, whereas analysis using other measures would suggest a much stronger pattern).

Of course, social statisticians are unlikely to be so naïve as to believe that the results of their analysis might not be influenced by other measurement strategies. Such appreciation has been built into the conduct of most social survey research at least since the early phase of critical responses to the survey method (cf. Cicourel, 1964 – although we should recognise that it is equally easy to find examples of the citation of social statistics with more deterministic interpretations than might reasonably be drawn if we recognise the possibility of alternative measurement instruments being used). Yet whilst most researchers fully appreciate the fallible nature of the construction of their variables, we nevertheless see very few examples of empirical research where researchers implement, review, discuss and document a selection of plausible measures; the more common model, by contrast, is to choose a single measurement device at an early stage of a research project, and use this measure throughout (perhaps, but not necessarily, preceded by a cursory theoretical discussion of potential options). The sequential nature of statistical analysis arguably contributes to these conservative tendencies: researchers are typically aware, for instance, that the inputs required at later stages of their work to

specify and run statistical models, or generate and interpret interaction effects between variables, for example, can be expected to expand linearly, and sometimes exponentially, as a function of the different available measurement options, thus providing a clear disincentive to comparing across multiple measures.

To change standards, it is clearly not sufficient to simply demonstrate the potential impact of different variable operationalisations. We instead speculate that to encourage a more scientific approach to generating survey statistics it may be necessary to actively reward that research which demonstrates good practice, and penalise that which does not. In contemporary academic and public dissemination research, it is arguable that the contrary prevails: communication and publication pressures preclude discussion of numerous different operationalisation details, whereas the time delays caused by comparing measurements thoroughly act to inhibit successful completion of a project.

The exciting opportunities presented by services such as the GESDE systems are that they may be able to take advantage of evolving technological opportunities to dramatically shift the cost-benefit equation in favour of a more rigorous, scientific approach to working with social survey data. Resources such as GESDE have the potential to dramatically reduce the time demands on researchers in obtaining information on, operationalising within their data, and suitably citing, a wider range of measurement options. The GESDE services are also positioned to publicise to the research community the range of measurement options which are easily considered, with the knock-on effect that analysis which deals adequately with all alternatives may be better valued, whilst analysis which neglects relevant options more critically received.

4. Conclusion

The GESDE services, developed for the UK ESRC-funded DAMES project, offer online interfaces to specialist social science data about measures of occupations, educational qualifications, and ethnicity. In all these areas it is well-known to researchers that multiple measurement strategies exist, but it is less common to observe social scientists dealing adequately with the range of options. The GESDE services are one of number of initiatives linked to the UK research councils' investment in e-Science which have sought to adapt technological developments to the benefit of scientific research in the domain of complex data resources (the MethodBox project. www.methodbox.org, at the University of Manchester also has many similar features to the GESDE services, whilst other information sharing projects such as MyExperiment, www.myexperiment.org, or PolicyGrid, e.g. Edwards et al. 2009, offer more generic resources which also have the potential to influence the statistical results of survey research). By responding to new technological initiatives, however, these projects face non-trivial challenges in balancing adequate provision for design and maintenance work, openness to new inputs, and the priorities of relevant user communities.

Acknowledgement

The GESDE projects are components of the ESRC funded 'Data Management through e-Social Science' research Node, RES-149-25-1066 (www.dames.org.uk).

References

- Birkin, M., Procter, R., Allan, R., Bechhofer, S., Buchan, I., Goble, C., et al. (2010). The elements of a computational infrastructure for social simulation. *Philosophical Transactions of the Royal Society, Series A*, 368(1925), 3797-3812.
- Buis, M. L. (2010). *Inequality of Educational Outcome and Inequality of Educational Opportunity in the Netherlands during the 20th Century*. Amsterdam: VU Univ. Am.
- Bulmer, H. (1956). Sociological analysis and the "variable". *American Sociological Review*, 21(6), 683-690.
- Chauvel, L. (2002). Educational Inequalities: Distribution of Knowledge, Social Origins and Social Outcomes. In Y. Lemel & H. H. Noll (Eds.), *Changing Structures of Inequality: A Comparative Perspective* (pp. 219-249). Montreal: McGill-Queens UP.
- Cicourel, A. V. (1964). *Method and Measurement in Sociology*. New York: Free Press.
- Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The Dependability of Behavioural Measures*. New York: Wiley.
- Dale, A. (2006). Quality Issues with Survey Research. *International Journal of Social Research Methodology*, 9(2), 143-158.
- Edwards, P., Farrington, J., Mellish, C., Phillip, L., Chorley, A., Heilkeme, F., et al. (2009). e-Social Science and Evidence-Based Policy Assessment: Challenges and Solutions. *Social Science Computer Review*, 27(4), 553-568.
- Freese, J. (2007). Replication Standards for Quantitative Social Science: Why Not Sociology? *Sociological Methods and Research*, 36(2), 153-171.
- Ganzeboom, H. G. B. (2011). Harry Ganzeboom's Tools for Deriving Status Measures. Retrieved 1 January, 2011, from <http://www.harryganzeboom.nl/isco08/>.
- Halfpenny, P., Procter, R., Lin, Y., & Voss, A. (2009). Developing the UK-based e-Social Science research programme. In N. Jankowski (Ed.), *e-Research: Transformation in Scholarly Practice*. London: Routledge.
- Heath, A. F., & Cheung, S. Y. (2007). The comparative study of ethnic minority disadvantage. In A. F. Heath & S. Y. Cheung (Eds.), *Unequal Chances: Ethnic Minorities in Western Labour Markets*. Oxford: Oxford University Press.
- Hoffmeyer-Zlotnik, J. H. P., & Warner, U. (2005). How to Measure Education in Cross-National Comparison. In J. H. P. Hoffmeyer-Zlotnik & J. Harkness (Eds.), *Methodological Aspects in Cross-National Research*. Mannheim: GESIS – ZUMA.
- Leiulfsrud, H., Bison, I., & Jensberg, H. (2005). *Social Class in Europe: European Social Survey 2002/3*. Trondheim: NTNU Samfunnsforskning/NTNU Social Research Ltd.
- Office for National Statistics. (2010). *Standard Occupational Classification 2010. Volume 3: The National Statistics Socio-economic Classification: (Rebased on the SOC2010) User Manual*. Basingstoke: Palgrave Macmillan.
- Rafferty, A., & Smith, S. (2008). *Variables, Datasets and Finding What You Want: Developing Online Search Tools*. Manchester: CCSR working paper 2008-11, Cathie Marsh Centre for Census and Survey Research, University of Manchester.
- Rose, D., & Harrison, E. (Eds.). (2010). *Social Class in Europe: An Introduction to the European Socio-economic Classification* London: Routledge.
- Schneider, S. L. (2008). *The International Standard Classification of Education (ISCED-97). An Evaluation of Content and Criterion Validity*. Mannheim: MZES.
- Schneider, S. L. (2010). Nominal comparability is not enough: (In-)Equivalence of construct validity of cross-national measures of educational attainment in the European Social Survey. *Research in Social Stratification and Mobility*.
- University of Essex, & Institute for Social and Economic Research. (2010). *British Household Panel Survey: Waves 1-18, 1991-2009 [computer file], 7th Edition*. Colchester, Essex: UK Data Archive [distributor], July 2010, SN: 5151.
- Vardigan, M., Heus, P., & Thomas, W. (2008). Data Documentation Initiative: Towards a Standard for the Social Sciences. *International J of Digital Curation*, 3(1), 107-113.