

ETHNICITY AND THE COMPARATIVE ANALYSIS OF CONTEMPORARY SURVEY DATA

Paul S. Lambert^{†*}
Stirling University

This paper discusses issues that arise in the cross-nationally comparative analysis of social survey data when interest concerns ‘ethnicity’ and related concepts. A number of both practical and theoretical problems arise, and a review of data in a selection of cross-nationally harmonised surveys reveals that current resources are a long way from satisfying the requirements of most social science analysts. However, it is argued that such problems are not an excuse for abandoning the comparative analysis of ethnic differences, and examples of alternative analytical strategies are reviewed for their adequacy.

Keywords: Ethnic group, nationality, religion, language, citizenship, cross-national surveys, category scoring.

†Department of Applied Social Science, Stirling University, Stirling, FK9 4LA, UK, paul.lambert@stirling.ac.uk . Paper presented to the 6th ISA RC33 conference, Amsterdam, August 17-20 2004, session “Harmonising Demographic and Socio-Economic Variables”.

*Acknowledgement: The data analysed in this paper was gratefully provided by: European Social Survey; UK Data Archive; Economic and Social Data Service (UK); Luxembourg Income Study; International Public Use Microdata Service. Any errors in the subsequent use of these datasets are the author’s responsibility. The research was assisted by an IRISS grant from the CEPS research institute.

Abstract:

Concepts of minority ethnic group are important to very many social science analyses, yet developments in their sociological conceptualisation generate serious problems for cross-national survey researchers. This occurs because as sociological discussions have expanded, ever more important qualitative differences between ethnic situations, in any particular country, have typically been propounded. In turn, the requirement of first mapping, then analysing, ethnic differences in terms of ‘comparable’ variables in cross-national quantitative datasets, is clearly rendered problematic. This challenge is not, however, one which social researchers should shy away from : history suggests that (comparative) social policy analysis on ethnicity, using whatever survey data categories are available, will continue apace, regardless of sociologists’ misgivings about the validity of the variable indicators available.

This paper explores the possibilities for analysis of ethnicity-related data on a number of contemporary cross-national studies (including the ESS, ISSP, LIS, WVS)[†]. The data considered ranges for instance across measures of religion, ethnic identity, nationality, country of origins and language spoken, and in the majority of cases, the data has been operationalised in different categorical forms for different national data sources. Two methodological strategies for comparative analysis are considered. The first position attempts to define a solution *a priori*, by seeking a strong model of analytical parsimony with regard to ethnicity differences that might be sensibly implemented on the large majority of relevant datasets. The second ‘relativist’ position examines the numerical indexing of different ethnicity categories within each country. Under this proposal, the index scores assigned to categories could simultaneously allow the incorporation of multiple ethnic differences in the analysis, along with a genuinely comparative evaluation of each ethnic location.

[†]European Social Survey; International Social Survey Programme; Luxembourg Income Study; World Values Survey.

1. Introduction

This paper discusses a number of issues that concern the analysis of ‘ethnicity’ and related social science concepts on the basis of social survey evidence. Those issues generate, almost unremittingly, problems and obstacles to the practitioner, the result of a variety of theoretical and pragmatic complications that are mentioned below. Moreover, this paper focuses on the desire to undertake comparative analyses using cross-nationally harmonised survey data collections, and it is little surprise to see that the problems of measurement and analysis of ethnicity are greater in this context than in national specific analysis. However, this review is driven by a pragmatic truism: that attempts at survey research on the topic of ethnicity will proceed apace, regardless of sociologist’s misgivings about the validity of the variable operationalisations considered. The pragmatic strategy, therefore, is to work with, rather than against, such interests and, as this paper attempts, to seek recommendations for ‘sociologically acceptable’ ways of incorporating ethnicity into cross-national survey data analyses.

The survey research method itself may be characterised as the systematic collection of the same pieces of information from multiple subjects; the operationalisation of that information in terms of numerically indexed variables; and the analysis of that information in terms of a summary of the patterns of relationships between those variables (Marsh 1982). In the social sciences, methodologists have generally had much to say about optimising each stage of the survey analysis process. Nevertheless, only a tiny fraction of the survey methods literature reflects upon issues of dealing with ethnicity. In contrast, leading cross-national methodological texts often avoid the subject completely (Merritt and Rokkan 1966; Hantrais and Mangen 1996; Harkness et al 2003). Alternatively, when attention is paid, it is very often restricted to an overwhelmingly critical perspective, whether within a cross-national framework (Allen and Macey 1990; Rea et al 1999; Favell 2003:25-34), or in national-specific literatures (for the example of the UK, see Fenton 1996; Ahmad 1999; Southworth 1999).

There are, however, a few evaluations of the comparative analysis of ethnicity through micro-social surveys (eg Lambert and Penn 2001; Hoffmeyer-Zlotnik 2003; Bonifazi and Strozza 2003). There are also various relevant outputs based upon cross-national survey research (eg Stille 1999; Heckmann et al 2001; Evans and Need 2002; SYSDM 2003; van Turbergen et al 2004; Jacobs and Tillie 2004¹). The following text reviews many of the issues in this field, as they relate to each element of the contemporary social survey process.

¹ Other ongoing projects in this field include the TIES project coordinated by M Crul (see <http://www.niwi.knaw.nl/en/oi/nod/onderzoeker/PRS1258604/toon>); and the “Ethnic minority disadvantage in the labour market project” coordinated by A Heath (see <http://www.britac.ac.uk/events/programmes/2003/031102emd-prog.html>).

1.1 Data Collection

Two issues of data collection have widely been commented upon. The first involves the survey interview process, a suggestion that the power dynamics of the social contact involved could bias, misdirect, or render ethically unsound, the collection of social survey evidence from members of minority ethnic or cultural groups (Back and Solomos 1993; Gunaratnam 2003). Though this critique has found influence with certain perspectives, it does not merit extensive attention for two reasons. Firstly, interview power dynamics represent a generic issue to social research methods which should not be regarded as uniquely important to the study of ethnicity. Secondly, extensive methodological guidelines exist designed to minimise power conflict issues, including prescriptions on the nature and form of questionnaire development and interviewer conduct (as for instance de Vaus 2002)².

The second data collection issue is more tangible and much more significant : ethnicity-based groups of interest usually comprise only small minorities of the national population, and random survey sampling methods face problems of low representation of cases from relevant minority groups. Some format of random sampling is essential to social survey design if there is any intention of inferential conclusions, yet the scale of most national sample surveys - typically a few thousand cases - tends to generate no more than tens of members of ethnic minority groups (see the next section of this review for ample illustrations). Moreover, problems of low representation are exacerbated by two further issues. In many countries, members of ethnic minority groups exhibit disproportionately higher non-response rates to social surveys (eg Simpson 1996). Additionally, in many countries, substantial diversity between ethnic minority subgroups is argued to be non-ignorable (cf Modood et al 1997), and thus a small initial subsample may need to be further divided into multiplicities of ever smaller and sparser groups of people³.

Problems of sampling representation are often unavoidable, as in the example of this review, when the relevant survey datasets were not collected primarily for the purpose of studying ethnicity differences. There are, however, some prospective strategies available. 'Oversamples' of relevant minority groups may be undertaken (contemporary examples include the German GSOEP and United States' PSID panel surveys, and Heckmann et al 2001; for methodological discussions of over-sampling techniques, see Brown and Ritchie 1981; Hughes et al 1995). Another common strategy is to maximise overall sample size, and thus increase small group representation, by using only the largest available data resources, or by pooling data from multiple sample surveys (eg Blackaby et al 1998; however, Heath et al 2001 describe instabilities which can limit such strategies). In the event of the multiplicity of ethnicity categories leading to sparsely represented categories, researchers often

² However, from the perspective of cross-national research it could be noted that a contemporary growth in support for pre-harmonised survey design (eg Harkness et al 2003; Lynn 2003) may be difficult to reconcile with the national specific sensitivities which conventional advice recommends in order to deal with such sensitivities.

³ This problematic can be further extended by claims to the non-ignorability of regionalised differences in ethnic group experiences (eg Green and Owen 1995).

choose to use their own judgement – carefully documented – to either merge the smallest groups with appropriate neighbours, or to exclude them from analysis altogether (see illustrations in the next sections). Lastly, in the event of disproportionate missing data, data-analytical and investigative strategies could be employed to account and control for disproportionate non-response by ethnicity (cf Jamshidian 2004).

1.2 Variable Operationalisations

It is the second element of the survey research process, that of ‘variable operationalisation’, which has occupied by far the most methodological attention with regard to the study of ethnicity. Hoffmeyer-Zlotnick (2003) offers a review of the issues for cross-national survey research. Firstly, Hoffmeyer-Zlotnick argues that social research in this field is complicated by lack of consistency in just which underlying concept it means to measure (see also Aspinall 2002). In previous literatures, the underlying concept of interest (or ‘referent’ as we use here, cf Lambert and Penn 2001), ranges over topics such as citizenship, national origins, country of ancestral origins, racialised visibility, language spoken, subjective cultural identity, and religion. In this writing, we use ‘ethnicity’ as an umbrella term to refer to any of these concepts, reflecting its use in British literature as a self-assigned category free to incorporate influences from any such referents (Banton 1997)⁴.

There then arises a question of how to adjudicate between alternative referents as sources of information for a survey based categorisation. Hoffmeyer-Zlotnick suggests a strategy for delineating the key dimensions of these referents in a manner amenable to social survey questions, leading to the prescription of a series of differentiations that all international surveys should seek to make (2003:p276). There have been many other procrastinations in this field, often featuring attempts to demonstrate the theoretical pre-eminence of certain or multiple referents as social science representations of ethnic differences (for cross-national perspectives, see Wrench and Solomos 1993; Lloyd 1995; Smith and Blanc 1995; Rea et al 1999; Aspinall 2002). However, very few such works offer useful, prescriptive advice to survey analysts.

Such *a priori* debates on the choice between alternative ethnic referents can be presented as an ‘absolutist’ approach to the measurement of ethnic difference. In cross-national research, whilst a variety of choices of referents are available, the chosen strategy of the majority of active researchers has tended to involve focussing upon only a single ethnic referent. Example applications include the focus on immigrant or citizenship status (van Tubergen 2004); concern with language minorities (Chiswick and Miller 1995); or the concentration on only certain, particularly distinctive, ethnic-cultural groups (Panayi 1999; Model et al 1999). Whilst such an approach offers an appealing conceptual clarity, it also does a

⁴ This term is not ideal, since in other circumstances, ‘ethnicity’ is taken to refer specifically to subjective cultural identity.

disservice to the tremendous complexity of most countries' ethnic mosaics, where situations cross-cut alternative referents in sociologically important ways. In section 3 below, we discuss the strengths and limitations of devising such absolutist ethnic schemes before analysis, with regard to contemporary survey evidence.

An important constraining feature of cross-national ethnicity research is that different nations tend to have their own preferences over the importance and measurement of alternative ethnic referents (Lambert and Penn 2001; Hoffmeyer-Zlotnik 2003; Favell 2003). Such national traditions are often strongly politicised and rigidly enforced – see for example Favell's (2001) description of the contrast between British and French views on the official recording of 'racialised' categories. Indeed, national institutions' influence over the research data available can cause considerable difficulties in obtaining comparable data – illustrated for example by attempts at coordinating reviews of labour market situations reported in two recent European Employment Observatory Reviews (Stille 1999; SYSDEM 2003).

Moreover, whilst national institutions often generate their own 'official' preferences for ethnic categorisations, this may be set against enormous difficulties within nations in achieving adequate consensus over *any* choice of ethnic categories. The British example makes an appropriate illustration – despite extensive consultation with social science practitioners during their development (Sillitoe and White 1992; Owen 1996), the 'subjective ethnic group' categories offered for the 1991 and 2001 OPCS national census's remain subject to vast swathes of critical reviews (Ballard 1997; Aspinall 2002, 2003). Indeed, political and ideological debates over appropriate categorisations often lead to the very forceful critique of schemas through the argument that they promote the (undesirable) reification of the very ethnic differences that they parameterise (Ballard 1997).

At the heart of this problem is an inherent tension between coding parsimony and theoretical complexity in the field of ethnicity research. On the one hand, a categorical schema attractive to survey research would generally have relatively few different values, and plenty of cases in each category. On the other hand, sociologists in most countries have a steady track record of producing literature emphasising the expansion of alternative ethnic referent categories, and, moreover, the temporal dynamism of whichever categories are favoured (Aspinall 2002). Indeed, in much of the field, the promotion and discussion of diversity of ethnic differences dominates contemporary discourse, and convincing sociological demonstrations of the importance of multiple categorical differences have led many writers to the conclusion that survey variable operationalisations can only offer at best a weak analytical device for representing ethnic diversity (eg Ahmad 1999).

Alternatively, it can be claimed that such perspectives are not useful, and fail to recognise the continuing desire for discussions of ethnicity differences through survey data records (and by implication, in the form of parsimonious categorical representations). The ideal situation may involve a social survey categorisation of sufficient complexity and sensitivity to capture the large bulk of relevant ethnic

diversities, but sufficient parsimony to allow communicable analysis and description; sections 3 and 4 below make suggestions in these terms.

1.3 Variable Analyses

Turning lastly to survey research as the analysis of data through the summary of relationships between variable representations, it could first be noted that so problematic are issues of data collection and operationalisation in comparative ethnicity research, that few methodological reviews of working with ethnicity data reach this stage of discussion. This is unfortunate, because in current research examples, many significant abuses of ethnicity data representations occur at the data analysis stage of the research process.

A first issue concerns the samples upon which cross-national analyses are conducted. One option is a 'pooled' cross-national data analysis (where cases from multiple countries are analysed simultaneously). Absolute comparability of categorical variable operationalisations is often seen as a pre-requisite for conducting this type of research - yet such classifications, by definition, are in danger of ignoring important national specific differentiations. An alternative is country-by-country analysis (the same analyses repeated in different countries then outputs compared). Technically, this format of analysis does not require comparability in variable specifications, although in many examples, absolute comparability of meaning is still generally regarded as beneficial for the communication of results. Subsequently, in many previous examples, whilst further ethnic differentiations have been available, researchers have often preferred to concentrate on only those limited elements of information that are comparably collected between countries.

A second issue concerns how categorical ethnicity variables contribute to the data analysis. Alternatives are influenced by the analytical intentions – whether ethnic boundaries mark key outcome differences that are to be explored in depth, or whether they simply contribute 'controlling' information to a wider analysis. Both applications are common, and to illustrate both in the analyses below, we give examples of tests of the employment/educational position of different ethnic categories (outcomes), and of regression models predicting economic attainment, with ethnic information as background predictors (controls). In either context, by conventional methods, the categorical ethnicity differentiation data is nominal in character and may only be operationalised in terms of 'dummy' variable factoring – that is, any individual category amounts to the 'unique' dichotomous variable of whether or not the case is a member of the category. In these situations, parsimony, and ideally dichotomy, in variable specifications, has strong data analytical attractions.

The key issue in data analysis of ethnicity information has traditionally concerned the pressure to 'collapse categories' – that is, to recode original categorical data into increasingly simple representations. This pressure arises for two reasons. Firstly, as noted above, the representation of ethnic minority cases in national sample surveys

tends to be low, and the distribution of cases between categories tends to be correspondingly sparse and skewed. Conventional wisdom in survey analysis argues that categories should hold as many cases as possible, and that sparse groups are of little use to research. Common solutions are to recode several minority categories into just one; to ignore the smallest categories altogether; or to merge together differentiations which could alternatively comprise distinct variables. Secondly, where data analysis involves the summarising of relationships between different variables, a consistent motivation for cognitive convenience is simply to limit the number of concepts involved in any given analysis. This generates a desire - driven solely by the nature of data analysis intentions - to limit ethnicity related variables to simple, parsimonious referents. Thus, particularly common in cross-national research is the deliberate restriction of attention to only one or two referents (most often 'immigration status') for the purposes of comparative clarity.

Whilst the provenance of both of these pressures is obvious, it is possible to argue that they can be overstated. Firstly, whilst the problem of sparsely represented categories is significant, the nature of the problem does depend upon the wider analytical question. For instance, if generalisations about distributional statistics for different ethnic categories are desired, then only large numbers of cases in each category are likely to sustain a reasonable level of confidence in the inferential significance of observed patterns. Similarly, if researchers are interested in studying differences in the structural processes experienced by different ethnicity groups, then a useful analytical device, a statistical model of the process outcome, requires the specification of multiple main and interaction effects with ethnicity groups, or the structural separation of models on each group, and such options will only be comfortably sustained when large numbers of cases represent each group. However, one common misconception is that a skewed distribution of cases is inherently problematic in such treatments, when in fact it is solely the absolute number of cases representing a minority group that is important. Moreover, there are certain types of analytical method where higher absolute representations are not so important – for instance if the ethnic differences are included more as a 'control' than outcome feature, or where inferential analytical techniques are not so important to the research.

Secondly, whilst the desire for conceptual clarity that encourages the limitation of data to only a few concepts has a clear communicative benefit, it is wrong to believe that techniques of data analysis inherently encourage such strategies. A myth to this effect apparently arises from the narrow range of data analysis techniques which social science researchers more commonly employ: 1-, 2- or 3-variable descriptive comparisons. In fact, many multivariate analytical techniques are available to social scientists which can readily incorporate a complex array of conceptual indicators and meaningfully summarise their interactions. Whilst there are dangers associated with utilising variable categories with limited numbers of cases (such as the risk of 'overfitting' data in statistical models), multivariate analyses nonetheless offer certain analytical treatments where certain forms of sparse data need not prevent the incorporation of complexity (such as discussed in section 4). Indeed, it can be claimed that a true 'optimum' minimum number of cases in any given category is never analytically appropriate - although many social scientists nevertheless operate with *de*

facto ideas about minimum sizes, arising from ceding an inappropriate primacy to univariate analytical methods⁵.

2. Review: Ethnicity and Comparative Survey Data

Described below are four examples of cross-nationally harmonised social survey datasets that are made freely available to secondary data analysts. Such datasets tend to have considerable, and often under-exploited, research potential in the social sciences (eg Kiecolt and Nathan 1985). We use the European Social Survey (ESS); the International Social Science Project datasets (ISSP); the World Values Survey datasets (WVS); and the studies of the Luxembourg Income Study (LIS). The ESS is a recent project conducting general attitude reviews of approximately 1000 adults per country, with most European countries involved in the supply of data for the first samples from 2002/3 (see Jowell 2003 and www.europeansocialsurvey.org/; the ESS data is accessed via the Norwegian Social Science Data Surveys). The ISSP has been established since the mid-1980's, coordinating the collection and distribution of sample surveys on selected themes each year from participating countries (see Braun and Uher 2003 and www.issp.org; the ISSP data was accessed via the UK Data Archive at the University of Essex). The WVS project has a longer history still, again involving relatively small probability samples in each member country, with collated questions on values focussing around a series of 'waves' of data collection (Inglehart 2000 and www.worldvaluessurvey.org/; the WVS data was accessed via the UK Data Archive at the University of Essex). The LIS project is an undertaking in the harmonisation of data from national probability samples such as Labour Force Surveys and general social surveys, extracting information relevant to the analysis of income and economic inequalities and scrutinising the data for comparability (see an earlier review as Lambert and Penn 2001, or www.lisproject.org; the LIS may be accessed direct via the project webpages).

The survey collections differ considerably in their design and data collection strategies. The most significant differences, following Harkness et al (2003), concern the data collection strategies. The ESS attempts to follow the highest standards of 'pre-harmonisation' of questions before going to field, resulting in high levels of complete questions across different countries. Both the WVS and ISSP similarly use pre-harmonisation techniques, although more flexibility between countries and time periods is built into their designs. Lastly the LIS surveys are entirely '*ex-post*' harmonised with one consequence being more examples of incomparable questions between countries, as well as higher risks of coding and translation errors (eg van Deth 2003).

⁵ It should be remembered that the problem of representation of sparse groups occurs as a function of combinations of relevant variables, not independently between variables – a group represented by 40 cases, say, might realistically only be thought of as four groups of 10, if cross-classification by gender and another dichotomous variable is relevant to analysis.

Table 1 reviews the contents of the most recent datasets for the relevant countries (gaps represent no data collection). The review broadly tests Hoffmeyer-Zlotnick's (2003) call for information on dimensions of ethnicity: the letters indicate whether information differentiating membership of the relevant categories can be found in the relevant surveys. Firstly, we see that differences in design have an evident impact: the ESS data have high very levels of consistency in their available resources, the ISSP and WVS studies are somewhere intermediate, whilst the *ex-post* data collections of LIS have considerably more variety of coverage.

We also see that the four groups of surveys tend to have slightly different themes of coverage. The ESS data appears to be the most complete, offering differentiating information on every concept checked for in the analysis. The WVS data is also apparently very thorough, usually covering every relevant concept with the exception of citizenship / nationality and parental origins. The ISSP on the other hand does not consistently offer much ethnicity data – only religious denomination is regularly recorded in most countries, then additionally, in various circumstances, measures of subjective ethnic group, language, or country of birth or citizenship may be available (in fact, the ISSP questionnaire allowed national organisations their own role in this question, and the variety reflects alternative national traditions). The LIS datasets also have low levels of detail on ethnicity data, measuring alternate referents in different countries (again reflecting the derivation of the harmonised data from national specific designs).

However, closer inspection of the data from any particular study quickly reveals the superficiality of many of the apparently complete records. Tables 2a to 2d look into each study's data in a little more depth, systematically charting four measures of the 'wealth' of the data across the surveys⁶. Almost all of the differentiations recorded are categorical in nature, and the first column [#Cat] indicates simply the number of unique categories measured by the relevant variables. The second and third columns then indicate the sparsity and skew of the relevant distributions. Skewness, as the proportion of cases clustering into the largest category [Skew], is a quick indicator of how much variation is likely to be usefully analysed – a highly skewed variable can offer little differentiating information between cases, especially if the overall sample size is relatively small. The sparsity measure [#NSC] checks the absolute number of cases in relevant ethnicity categories - by our measure, the number of categories with more than 50 or 100 people representing them. This again indicates how much analysis can realistically be undertaken on the variables, and the absolute number of cases is ultimately more important than skew in this regard. Lastly, the fourth column [%m] shows the number of missing cases for each relevant variable – in some literatures it is generally expected that questions relating to ethnicity will be characterised by high levels of missing data (due to refusals), though this is not generally borne out across the range of datasets shown in tables 2a to d. Overall, tables 2a to d, in conjunction with a look through the specific distributions of contributory countries and the relevant survey documentation details, do not give encouraging evidence of high data 'wealth' for studying ethnicity differences.

⁶ We use the term 'wealth' in reference to survey analysts' common descriptions of the 'richness of the data': wealthy data contains high variability across cases on the relevant variables, and thus 'relationships between variables have the maximum chance to show up' (Punch 2003:p38).

At first glance, the ESS seemed the most impressive of the survey collections. However, Table 2a reveals a clear pattern of an 'impoverished' data resource on ethnicity for the majority of countries: although missing data is minimal, the data distributions are highly sparse and skewed, and with small numbers of cases already representing each country, there are in practice likely to be very few circumstances where the ESS divisions will sustain an informative analysis of ethnicity. Compounding the distributional problem, closer inspection of the ESS variables reveals that the completeness of the ESS data has sometimes been achieved at the cost of some of the more theoretically appealing features of ethnic difference. The many categories of country of birth, citizenship and language are valuable. However there is no similar level of detail indicating ethnic identity divisions (simply a dichotomous record of whether or not the subject identifies with any minority group). Equally the measures of parental place birth (collected for both mother and father, allowing us to discern small numbers of 'mixed' parentage cases), only differentiate 5 alternate continents. Thus, within any given country, the ESS data may well not be able to engage with mainstream debates over ethnic identities.

Table 2b shows firstly that the ISSP data also shares similar problems of sparsity and skewness with the ESS. In a few countries, for instance Canada, Latvia and the USA, the spread of cases by ethnicity categories is wider, whilst the religious variable measures also often show moderate differentiation. Generally however, the small number of cases per country, combined with the skew of most variables, is likely to severely curtail the value of ISSP analyses of ethnic minority situations. The bulk of the relevant ISSP data concerns religious denomination, and inspection of the ISSP documentation and variable distributions reveals a problem which is also shared with data from the ESS and WVS. This is that, whilst the ISSP offers consistent and moderately 'wealthy' differentiations by religious denomination, in practice the variety of categories represented does not engage with the religious divisions of interest to most researchers of ethnic identities. The well represented categories are in nearly all cases simply the division between 'no religion' and 'Christian', with perhaps one or two further divisions by major Christian denominations (the main exception to this rule being Israel, where the religious measure taps directly into the main ethnic division of most popular interest). By contrast, most social scientists are more interested in more visibly different religious differentiations such as those involving Asian or Arabic religions (cf Modood 1991). Whilst the ISSP and other surveys do measure these categories (somewhat in contrast to the more pessimistic review of Brown 2000), the number of cases involved is consistently sparse.

The WVS data again shows similar features to the ESS and ISSP : Table 2c suggests that for most countries, the extensive information offered by the WVS is severely compromised by the skew and sparsity of the data distributions. Again however, there are a number of WVS countries where a moderate degree of differentiation by ethnicity category is evidence, and the WVS is perhaps particularly valuable for its inclusion of data from countries where ethnic differences are widely politicised but seldom accessible to survey data, such as those of the Balkans and Eastern Europe (in fact the WVS wave 3 included many further countries outwith Europe and North

America which are not mentioned here for convenience). A drawback in working with the WVS data on ethnicity, however, is that the level of documentation in the centrally distributed files can be somewhat limited: readers must cross-check between multiple documents to obtain pithy descriptions of category labels, and, in a few example countries, apparent ethnic group data has no accessible documentation pertaining to it (thus excluded from the tables shown here).

Lastly, the LIS data studies have several differences from the alternative collections. Although Table 2d reveals they still have similar levels of skew in the relevant distributions, it can be noted that the number of cases in most LIS surveys is much larger than other studies (though there is considerable variety here). This means that, crucially, sparsity of representation of categories is often less of a problem. Indeed, reviewing largely the same data, Lambert and Penn (2001) were able to highlight a number of countries where ethnicity related data could reasonably be used for meaningful analyses. As mentioned previously, a characteristic of the LIS data in particular is the wide range of ethnic referents used in constructing the relevant variables, and whilst this national specific influence makes for a conceptual headache for comparative analysis, it does simultaneously improve the ‘wealth’ of the LIS data and increase the levels to which its particular categories engage with national specific research concerns (for instance, ‘ethnic self-identity’ in Britain; citizenship in Germany).

By most standards, the picture of ethnicity information available on these four groups of surveys is messy and problematic. It should be remembered, moreover, that harmonised survey resources like these represent the stronger examples of comparative survey resources. Many other social researchers have attempted to conduct comparative analysis involving ethnicity on other survey data collections which have not been subjected to the same levels of harmonisation and documentation (for instance, van Tubergen 2004; Stille 1999 describes investigations using contemporary labour force survey data). In the next two sections, we describe some of the possibilities of analysis that are open to us with such problematic data collections.

3. Prescriptions (1): the assertion of categorical boundaries

The large majority of survey analysts have operationalised ethnicity information in terms of categorical variables. These offer clear advantages of specification and communication, since in most cases it is relatively easy to document the criteria or differentiation that separates categories. However, as we have seen, a wide literature has been critical of such approaches, chiefly through arguments about the appropriate choice of category definitions and the analytical usefulness of the derived units. Within countries, an ‘official’ advocacy of a set of definitions thought to most fully capture the key elements of ethnic inequality sit uneasily with protracted sociological debates, and offer little absolute comparability to cross-national researchers, who instead often concentrate on relatively pithy ethnic categorisations.

Below we evaluate a few alternatives relevant to the harmonised datasets considered here. Broadly following Hoffmeyer-Zlotnick's (2003) suggestions, listed below are a selective choice of six categorical models for representing ethnic differences in cross-nationally comparative surveys:

- 1) **[IMM] Immigrant status:** a dichotomy indicating whether or not a case was born in the current country. This measure is very widely used in previous research (eg van Turbegen 2004). However it is very limited in the information it conveys about ethnic differences – for instance it fails to make internal differentiations between different immigrant backgrounds, and it cannot recognise non-immigrant minorities, or even the children or grandchildren of immigrant ancestors (eg Banton 1997). It is also flawed because of its popular conflation with concepts of citizenship - Hoffmeyer-Zlotnik (2003) for instance emphasises how the two measures are analytically distinct).
- 2) **[LAN] Minority language use:** a dichotomy indicating whether or not a case generally speaks a language other than an official majority language of the host country. This measure has considerable sociological significance, as the analysis of ethnic differences is increasingly concerned with information on language use (eg Portes and Rumbaut 2001; Alba and Nee 2003). However this measure has several flaws: there is a lack of cross-national consistency to the measurement of 'minority' languages, and a subjectivity to individuals' reports of usage. It is also prone to ignore members of minority groups who do not use minority languages
- 3) **[VIS] Visible minority group status:** a dichotomy indicating whether or not a case belongs to a minority group on the grounds of any overt ethnic group formation – such as racialised visibility, subjective ethnic group identity, or participation in a 'visible' minority religion⁷. This measure highlights the ethnicity referents which are most popular with leading sociological thinking in the field (cf Modood 2002); however, it painfully conceals any further diversity and differentiation.
- 4) **[MIN] Any minority group membership:** a dichotomy which extends the VIS categorisation to highlight any ethnic minority identity from any relevant referent – citizenship, country of birth, parental national origins, language use, and visible minority group status. Again, this catholic measure would seem highly problematic for its masking of diversity within its categories. It is also very unlikely that the same differentiations will contribute to the same categories in different countries.
- 5) **[CON] National-specific scheme:** a multiple categorical scheme chosen subjectively from the available schemas favoured by the national literature. This has clear substantive attractions, but is problematic for comparative purposes.
- 6) **[EC9] Comparative ethnicity 9 category measure:** a multiple categorical scheme advocated here as an attempt at incorporating the most important ethnicity differentiations in an absolutist cross-nationally comparable way (cf Hoffmeyer-Zlotnick 2003). It uses information from five dichotomous measures, IMM, LAN, VIS, and indicators of citizenship and of whether or not either parent was born outside the country. The information combines in the categories listed below, where 'minority group' corresponds to either VIS=1, or having non-host citizenship:

1 CCNN – Self and parents born in country, no minority group or language

2 CCMN – Self and parents born in country, minority group but not language

⁷ This latter criteria regarding religion is constructed in this analysis. It reflects the perspective (and approximation), advocated for instance by Modood (1991) and Brown (2000) that only certain minority religions are most important to social stratification outcomes in the contemporary societies under study, and here is defined as recording of a South Asian or Islamic religious identity.

- 3 CCL– Self and parents born in country, minority language
- 4 CPNN – Self born in country, parents not, no minority group or language
- 5 CPMN – Self born in country, parents not, minority group but not language
- 6 CPL – Self born in country, parents abroad, minority language
- 7 FNN – Self born abroad, no minority group or language
- 8 FMN – Self born abroad, minority group but not language
- 9 FL – Self born abroad, minority language

Table 3 summarises the distribution of cases to these categorical schemes from the four survey collections considered here. The first two columns show the distribution of cases in the pooled (cross-country) datasets from the ESS (a pooled dataset from these two surveys is much more readily constructed than for the ISSP and LIS studies). This somewhat artificial aggregation merely serves to illustrate overall patterns of data availability and typical distributions of cases. We can note from these columns the difficulty of defining a coherent ‘CON’ scheme on a pooled cross-national datasets; and that, whilst the skewness of most minority categorisations persists, the heterogenous ‘minority’ categorisation does involve a substantial proportion of cases.

National specific evaluations are clearly more informative, and in Table 3 we use the second two sets of columns to illustrate data availability across the surveys for the examples of the UK and Germany. We see clear differences in the availability of different measures across different surveys. In particular, our ‘optimum’ EC9 measure can only be derived from the ESS (the only survey which adequately makes all of the relevant referent differentiations), whilst outwith the ESS data, only a minority of differentiations can be satisfactorily made for any given survey sample. On the other hand, the generally rich ESS data lacks, for the example of Britain, adequate specification of the national specific ‘CON’ scheme (here taken as the OPCS census classification). For both the UK and Germany, we also see that the majority of dichotomous differentiations, when available, do not suffer excessively from the sparse representation of cases, though for all but the larger LIS studies, the ‘CON’ schemes do not have substantial numbers of cases in their minority categories. The EC9 categories, however, largely avoid sparsity problems, with the clear exception of the CCL and CPL categories (minority groups with regard to language but not other criteria).

Table 4 then evaluates associations between the various categorical measures considered in Table 3, and a variety of measures of demographic and social stratificational differences that might be expected to have some relationships with ethnicity differences, using the 2002 ESS data. The first rows of each of the three panels in the table show coefficients which illustrate the magnitude of bivariate associations between each alternative measure of ethnic difference, and assorted measures of age; educational experience; recent occupational advantage (using the ISEI socio-economic status scale of Ganzeboom and Treiman 1996, for current or past employment); a measure indicating level of use of the internet (a likert scale treated as metric); and a measure indicating self-placement of political views on a

left-right scale (also a likert scale treated as a metric)⁸. These sets of associations indicate several interesting trends. Generally speaking, associations with all sets of indicators are discernible but modest, and all follow the same patterns of difference. Measures of language status are more strongly associated with most concepts that are country of birth measures. Associations with the larger, more heterogeneous ‘MIN’ group are generally larger in magnitude than those involving more precise concepts, with the exception of educational and occupational differences in Germany. However, it is the categorical schemes that extend beyond dichotomous measures that generally show the strongest patterns of association, with both the CON and EC9 patterns representing significant jumps in explanation. Unexpectedly, the EC9 measures generally capture slightly more association than the national specific schemes – suggesting that their multidimensionality effectively captures a substantial proportion of ethnic differences which are ignored by the politically expedient national specific measures.

The bivariate associations reported in Table 4 are intended to represent the magnitude to which the relevant ethnic differentiations associate with alternative factors: if we accept that such factors tend to genuinely differ by ethnicity circumstances, we can argue that the larger the association, the more effective the analytical representation. Additionally, the last rows in each panel of Table 4 then consider the role of ethnic category dummy variables as ‘control’ factors in a regression model predicting occupational status (ISEI) differences – the expectation again being that the more explanation the ethnic category dummies can add to the model as main effects, the better they represent genuine ethnic differences. The slight variation in regression R² values and the significance of dummy variable effects suggests that ethnic differences do make a small difference to the regression model output and that different modes of representing them have alternative impacts. However, the lack of significance of many of the dummy variable effects is ambiguous: it could indicate a genuine absence of an ethnic impact on the overall outcome; however, since previous research findings would suggest otherwise, it is more likely to indicate the inability of the regression models to identify genuine patterns of association with confidence, most likely a function of the low number of cases representing certain categories. Most important, as shown in the last rows of each panel, it can also be noted that the alternative ethnic controls for the relevant models did often have slight impacts on the interpretation of relative effects of other variables in the equation⁹. This suggests that the choice and operationalisation of categories is influential on this outcome

This evaluation of categorical representations of ethnic differences helps illustrates several issues. Firstly, a number of dichotomous measures of concepts are readily operationalised in cross-national surveys and exhibit discernible associations with other relevant factors; they may also be easily understood and communicated. They are, however, compromised by ambiguity as a result of a capacity for substantial

⁸ Whilst several of these measures represent rather crude variable indicators of the concepts involved, they may still be expected to exhibit the same properties of differential association as a more precisely defined indicator.

⁹ Interpreted as the non-overlap of 95% confidence intervals on the main effect parameter estimates. Further evaluation of the significance of interaction effects between ethnic categories and other main effect variables would also be useful to this evaluation.

internal heterogeneity (according to further ethnic differences), and this may mask precisely the differences that social scientists may most be interested in. Alternatively, national specific analyses of 'CON' variables are seriously constrained, by the impact of a lack of conceptual comparability between countries, by the lack of availability of the relevant data for many surveys, and by the sparse representation of certain contributing categories. Lastly, the 'EC9' categorical representation has both conceptual attractions (as a consistent measure between countries) and weaknesses (its possible disengagement from national-specific concerns over ethnicity differences). As a practical measure the EC9 categorisation has intermediate properties. Its association with other variables is particularly strong, and in most circumstances there is a moderate spread of cases between different categories. However, categories of the measure still tend towards sparsity in certain contexts, whilst many of its categories continue to conflate sociologically significant ethnicity differences. It may thus be argued that with harmonised social surveys, it remains very difficult, if not impossible, to devise and analyse satisfactory categorical ethnicity classifications which apply to national populations in a comprehensive but meaningful way.

4. Prescriptions(2): Ethnic category scoring

The preceding writing has discussed – and in large measure problematised – the analysis of ethnicity differences in social science surveys through the medium of categorisations of ethnicity differences. This section asks whether there is any alternative treatment of ethnicity data that is available for survey analysts. It argues that it can be defensible to collect categorical data in terms of multiple differences between ethnic referents, but then analyse that data by assigning metric scores to different categorical levels (for earlier developments of this argument, see Lambert and Penn 2001, and Lambert 2002). This could have the advantage of bypassing the pragmatic difficulties associated with the low representation of many minority group categories. Additionally, if the assignment of scores is done in a consistent way, it is also possible that the assignment process itself could add to the cross-national comparability of the variable operationalisations.

A first relevant point is to recognise that metric treatments of categorical data on ethnicity are already extant. In economics for instance, the use of variables for 'time since immigration' is often used as a proxy for the magnitude of ethnic differences. More trivially, any dummy variable representation of ethnic category differences can ultimately be presented as a metric contrast. Moreover, many analyses of ethnic differences already resort to the metric scoring or ranking of categories as a summary cognitive device, whilst in a few other applications, category scores have also been derived as indicators of ethnic differences (cf Prandy 1979).

Table 5 demonstrates how one such derivation of ethnic category scores may be obtained. The columns of Table 5 show the results from a series of Stereotyped

Ordered Regression (SOR) models (Anderson 1984). Here, ethnic category scores are derived as a function of average differences between ethnic categories according to a variety of predictor variables (the SOR models were derived using macros following Hendrickx 2000; for an extended discussion of the derivation of these models for this context, see Lambert 2002). The SOR scores derived for a selection of EC9 and CON categories are shown in the middle panel: the values represent a dimension of difference between categories as it is reflected in typical differences in the predictor variables in the pattern of the coefficient estimates shown in the first pattern. For instance, the EC9 category 6 SOR scores are generally high positive (for the category born in this country but parents born abroad and speaking a foreign language). This indicates on average higher values on the predictor variables with positive coefficients (such as educational level and age) but lower average values on the predictor variables with negative coefficients (such as ISEI score and current employment). The idea behind this derivation is that the SOR scores represent relative differences between ethnic categories in a country along a dimension of social stratificational inequality, since the predictors involved in the estimation reflect typical markers of social stratification experiences.

In this case, a measure is thus derived which reflects *relative* ethnic difference with each country. Alternative SOR score derivations could be achieved by changing the selection of predictor variables, or indeed the number of categories for which a score is derived (some alternative SOR score derivations, including one which involves a summary of life history information, are described in Lambert 2002). Indeed, in principle, SOR scores could be estimated for the dozens of different categories which emerge when cross-classifying multiple ethnic referents, and they could be estimated with slightly different functions in different countries if it is felt to be appropriate.

Additionally, there are of course many other ways in which ethnic categories could reasonably be assigned score values. For instance this could be through using other summarising functions which have perhaps simpler and more communicable interpretations (say the average employment advantage score associated with each group), or perhaps by *fiat*, allowing sociological thinkers an opportunity to make judgements reflecting a wider range of considerations concerning ethnic differences. However derived, category scoring offers an opportunity to analyse ethnicity differences in hierarchical manners, interpreted as the typical effects of ethnicity differences as they operate through a given structure of inequalities.

It is also important to highlight that metric variable representations need not be conflated with uni-dimensionality. The examples give in Table 5 all involve just one dimension of score structures. However there is no reason why further orders of scores or further dummy variable contrasts cannot also be added to incorporate information on more than one feature (this idea is explored briefly in Lambert and Penn 2001 and Lambert 2002, although the SORs scores used in those examples are predominantly illustrations of one dimension of differences).

It can then be argued that such relative measures of ethnic differences have a form of cross-national comparability, in the same fashion say as a measure of income that is standardised around national averages. By contrast to absolutist measures of ethnic category differences, such relative differences could have a consistent interpretation reflecting the most prominent ethnic structures within each country considered. They would also have the advantage, as metric variables rather than categories, of being amenable to forms of analysis that are more problematic with traditional categorical variable representations, since the sparse representation of certain categories is not as problematic for inferential analyses. Moreover, any choice of ethnic category scores would clearly not purport to be fixed in time, and changing choices in the scoring of categories could thus be taken as a way of dealing with dynamism of important ethnic differences over time.

The potential value of such ethnic category scoring is that it may parsimoniously summarise the key elements of ethnic category differences that affect the population under study. In the last panel of Table 5 however, we see only mixed evidence on this issue, at least for the examples illustrated. The derived scores do indeed exhibit patterns of association with other variables of a nature that would be expected from the analysis of other ethnic categorical information. However, in the last rows summarising the output from regression models, we see that they have almost no influence as control variables in predictive regression models. This latter finding corresponds with those of Lambert and Penn (2001) and Lambert (2002), who found only a few circumstances where SOR effects altered the estimates from predictive regression models. However, the summaries of Table 5 only introduce a few examples of SOR score effects; there is a strong possibility, also discussed in the preceding articles, that a more powerful relativist scoring framework could be uncovered, and thus that the assignment of ethnic category scores could prove a significant benefit to the cross-national analysis.

5. Conclusions

Large scale social survey analysis using existing secondary data has great potential for social science investigations - some authors have argued it represents the research method that most satisfactorily allows for the formulation and testing of theories on the basis of research evidence, whilst retaining some confidence in wider generalisability (eg Goldthorpe 2000; Steuer 2003). However, existing cross-national survey resources typically have weak and highly problematic information on ethnicity differences. Whilst this evidence is still often used for influential social policy analyses of ethnicity differences, more rigorous researchers more often become stuck or discouraged with the variety of problems which arise.

This paper has sought to outline some of the key areas of debate, structured around the topics of data collection, variable operationalisation and data analysis. Secondary survey analysts can only readily influence the latter two elements, and it is suggested

above that greater cross-national consistency in data collection and harmonisation can lead to more comparable ethnicity categories over time - although such categorical measures remain problematic as a result of their distributional and informational forms. It is also argued that a further variable operationalisation – namely, category scoring – can be used as a solution to issues of both sparse representation and the absolute comparability of the meaning of ethnic differences. This occurs because in principle the metric variates generated are not inhibited to the same degree by sparsely represented categories, and because it is argued that reference to a national specific derivation of relative location is the most appropriate form of cross-national analysis. Needless to say, the representation of ethnic category differences through score values still lacks convincing displays of its empirical superiority – though the possibilities remain open.

Table 1: Data availability by country and study

C Citizen of which country **P** Parental country of birth **E** Ethnic self-identity
B Country of birth – which **L** Which language used **R** Religious denomination
T Time in this country n No relevant data

Review conducted June 2004. Lower case letters when categories are dichotomy only – eg, born in host country or not is ‘b’ rather than ‘B’. Blank cells for non-coverage of country and data

	ESS	ISSP	WVS	LIS		ESS	ISSP	WVS	LIS
Australia		BR	BTLER	BT	Latvia		CLR	BTLER	
Austria	CBTPLeR	R		Cb	Lithu.			BTLER	
Belgium	CBTPLeR			C	Luxem.	CBTPLeR			CT
Bosnia			BTR		Maced.			bTR	
Bulgaria		RE	BTR		New Z.		ER		
Canada		LER		bT	Nthlds	CBTPLeR	R		
Croatia			BTER		N. Irel.		ER		
Cyprus		ER			Norway	CBTPLeR	R	BTR	Bp
Czech R	CBTPLeR	bR		C	Poland	CBTPLeR	BR	bER	n
Denmark	CBTPLeR	BR		CT	Portugal	CBTPLeR	cR		
Estonia			BTLER	E	Russia		ER	BTLER	BE
Finland	CBTPLeR	LR	bTLR	L	Serbia+M			bTLER	
France		R		C	Slovenia	CBTPLeR	ER	bTIER	n
Germany	CBTPLeR	CR	RBTE	CB	Spain	CBTPLeR	R	BTLER	n
Greece	CBTPLeR				Sweden	CBTPLeR	cR	bTLR	CBTp
Hungary	CBTPLeR	ER		E	Switz	CBTPLeR	CLR	BTLR	c
Ireland	CBTPLeR	R		CBT	UK (GB)	CBTPLeR	ER	E	E
Israel	CBTPLeR	BR		T,R-B	USA		ER	BTLER	cE
Italy	CBTPLeR	R		B					

ESS: all studies from 2002.

ISSP : all studies for 2000, except Australia, Cyprus, France, Hungary, Latvia, Poland (1999) and Italy (1998).

WVS : Wave 3 only 1995-7 (other countries covered by WVS in earlier waves and not W3). In some WVS countries, relevant data is nominally present, but all categories undocumented and listed here as missing.

LIS : uses latest available LIS study, all in range 1994-2001.

Table 2a: ESS data: Country and measure by data wealth

(Country and data codes as Table 1)

	#Cat	#NSC	Skew	%m		#Cat	#NSC	Skew	%m
	#Cat								
	#NSC								
	Skew								
	%m								
	ESS 2002: Typical sample size = 2100 cases per country.								
	Full ESS dataset					Israel, n=2499			
					C	6	1	98	2
C	118	5	96	0	B	60	5	66	1
	Country of citizenship				T	6	5	65	2
B	158	12	90	0	L	23	3	62	0
	Country of birth				E	2	2	87	5
T	6	5	90	0	P	7	4	34	1
	Time living in country (categories)				R	5	5	57	1
L	102	9	95	1					
	Language spoken at home				Norway, n=2036				
E	2	2	94	2	C	30	1	97	0
	Whether in a minority ethnic group				B	44	1	94	0
P	7	6	84	0	T	6	2	94	0
	Parents national origins (continent of parents' birth / mixed parentage)				L	23	1	97	0
R	9	9	40	1	E	2	1	98	0
	Religious denom. (current or past)				P	7	2	92	0
					R	9	4	51	0
	Data from selected other countries								
					Switzerland, n=2040				
C	24	1	95	0	C	41	2	89	0
B	36	1	92	0	B	63	3	83	0
T	6	2	92	0	T	5	3	83	0
L	17	2	63	0	L	32	4	63	0
E	2	1	98	2	E	2	2	95	0
P	6	2	84	1	P	7	2	71	0
R	9	2	61	1	R	9	4	38	1
					UK (GB), n=2052				
C	43	1	96	0	C	28	1	97	0
B	51	1	93	0	B	57	1	91	0
T	6	2	93	0	T	6	2	91	0
L	20	1	96	0	L	31	1	96	0
E	2	2	96	0	E	2	1	94	0
P	6	3	86	0	P	7	3	85	1
R	9	5	38	1	R	9	4	45	0

Table 2b: ISSP data: Country and measure by data wealth

(Country and data codes as table 1)

#Cat Number of categories in original data
#NSC Number of non-sparse categories (more than 50 cases, absolute value)
Skew Skewness: Percent of valid cases in the largest category
%m Percent of cases with missing data
ISSP: Typical sample size= 1000 cases per country.

	#Cat	#NSC	Skew	%m		#Cat	#NSC	Skew	%m
Asrl - B	42	2	80	4	Lat - C	3	3	59	0
Asrl - R	19	5	29	4	Lat - L	2	2	61	0
Astr - R	5	3	78	1	Lat - R	9	4	33	1
Bul - E	5	2	88	0	NZ - E	11	3	80	0
Bul - R	7	3	78	1	NZ - R	22	5	28	2
Can - L	4	2	66	0	Nir - E	5	1	99	0
Can - E	17	6	38	1	Nir - R	19	4	30	1
Can - R	27	3	53	25	Nth - R	7	3	64	0
Cyp - E	4	1	99	0	Nor - R	8	2	84	1
Cyp - R	2	1	99	0	Pol - B	5	1	97	0
CzR - b	2	1	95	3	Pol - R	6	2	91	1
CzR - R	6	2	58	3	Por - c	2	1	99	0
Den - B	9	1	95	3	Por - R	7	1	92	1
Den - R	5	2	85	3	Rus - E	6	3	84	1
Fin - L	3	2	93	1	Rus - R	12	4	60	4
Fin - R	5	2	85	1	Sve - E	7	1	91	0
Fra - R	8	2	60	0	Sve - R	7	2	71	4
Ger - C	13	1	94	0	Spn - R	6	2	87	1
Ger - R	7	3	44	0	Swe - c	3	2	87	1
Hun - E	7	1	94	0	Swe - R	5	2	62	1
Hun - R	10	3	52	0	Swi - C	16	1	88	0
Ire - R	10	2	83	0	Swi - L	10	3	58	0
Isr - B	38	3	62	0	Swi - R	9	3	43	1
Isr - R	2	2	87	0	UK - E	11	1	94	0
Ity - R	9	2	88	2	UK - R	14	4	42	0
					US - E	35	6	13	21
					US - R	15	7	24	0

NZ - E : categories are not mutually exclusive; Asrl - B : combines intra-Australian birth location data;
US - E : doesn't use conventional census ethnic identity categories, but 'country of ancestral origins';

Table 2c: WVS data: Country and measure by data wealth

(Country and data codes as table 1)

#Cat Number of categories in original data
#NSC Number of non-sparse categories (more than 50 cases, absolute value)
Skew Skewness: Percent of valid cases in the largest category
%m Percent of cases with missing data
WVS 1995-7: Typical sample size=1000 cases per country.

	#Cat	#NSC	Skew	%m		#Cat	#NSC	Skew	%m
<i>Full WVS wave 3 dataset 1995-7</i>					Finland, n=987				
					b	6	1	98	0
					T	5	1	98	0
B	100est	12	91	2	L	5	1	96	1
					R	6	2	81	1
T	7	7	92	21	Poland, n=1153				
L	100est	12	95est	9	E	5	1	97	0
					R	5	1	94	0
E	200est	10	90est	14	Spain, n=1211				
					B	8	1	97	1
R	14	11	38	4	T	6	1	98	1
<i>Data from selected other countries</i>					L	6	2	83	1
					E	5	5	44	2
					R	7	2	83	1
					Australia, n=2048				
B	7	4	78	0					
T	6	4	78	0					
L	5	1	99	5					
E	7	2	89	0					
R	10	3	48	1					
					Estonia, n=1021				
B	7	2	71	0					
T	7	3	71	1					
L	4	2	58	0					
E	6	3	54	1					
R	8	3	71	1					
Category estimates for total sample reflect non-harmonisation of documentation value labels									

Table 2d: LIS data: Country and measure by data wealth

(Country and data codes as table 1; two digits beside country name for LIS survey year)

#Cat Number of categories in original data
#NSC Number of non-sparse categories (more than 100 cases, absolute value)
Skew Skewness: Percent of valid cases in the largest category
%m Percent of cases with missing data
 LIS: Typical sample size= 50,000 cases per country

	#Cat	#NSC	Skew	%m		#Cat	#NSC	Skew	%m
Asrl94 – B	10	7	74	0	Lux00 – C	5	3	64	0
Asrl94 – T	9	5	74	0	Lux00 – T	m	m	65	0
Astr97 – C	2	2	90	3	Nor00 – B	98	6	93	0
Astr97 – b	31	1	95	3	Nor00 – p	7	5	90	0
Bel97 – C	5	2	74	22	Rom97 – E	5	5	90	0
Can00 – b	2	2	76	74	Rus00 – B	52	4	83	0
Can00 – T	5	5	25	72	Rus00 – E	16	3	92	0
CzR96 – C	9	2	99	7	Swe00 – C	81	2	96	0
Den97 – C	78	1	96	0	Swe00 – B	109	11	88	0
Den97 – T	15	4	94	0	Swe00 – T	m	m	87	0
Est00 – C	6	2	92	10	Swe00 – p	9	7	58	0
Fin00 – L	2	2	94	1	Swi92 – c	3	3	85	1
Fra94 – C	30	5	91	18	UK99 – E	9	7	95	5
Hun94 – E	2	2	97	10	US00 – c	5	5	84	0
Ger00 – C	68	5	91	0	US00 – E	5	5	70	0
Ger00 – B	67	5	94	3					
Ire96 – C	9	1	98	5	<i>m = 'metric' variable (time in years)</i>				
Ire96 – B	17	3	94	5					
Ire96 – T	m	m	95	5					
Isr01 – T	4	4	44	0					
Isr01 – RB	m	m	60	0					
Ity00 – b	2	2	98	0					

Isr-RP : religion (2-category) and birthplace inextricably conflated;

Table 3: Data distribution for selected categorical ethnicity measures

Absolute numbers, unweighted.

	Pooled studies		UK				Germany			
	ESS	WVS	ESS	ISSP	WVS	LIS	ESS	ISSP	WVS	LIS
IMM: Whether or not born in host country										
0	36839	21736	1860	n/a	n/a	n/a	2705	n/a	939	19732
1	3935	2129	191				214		34	1216
LAN: Whether or not speaks a minority language										
0	36383	20311	1978	n/a	n/a	n/a	2813	n/a	973	n/a
1	4336	3516	73				104		44	
VIS: Whether member of a visible minority group (ethnic identity, racialised category, religion)										
0	36261	16655	1865	913	1053	41756	2735	n/a	995	n/a
1	3655	5236	179	59	40	2351	161		55	
MIN: Whether or not a minority category member by any differentiation from Table 1										
0	29176	14015	1642	n/a	n/a	n/a	2432	n/a	929	n/a
1	10716	6886	393				457		55	
CON: National specific ethnicity categories										
1				1	8	2205	7	2	35	0
2	n/a	n/a	n/a	919	1044	41756	2799	1417	961	19669
3				12	12	387	26	20	7	719
4				7	4	192	38	10	2	311
5				9	2	76	49	15	2	103
6				5	11	535		37	10	879
7				2	7	418				
8				1	1	155				
9				3	3	98				
10				13	1	490				
EC9: Minority category membership by birth and parental birth, visibility and language use										
1 – CCNN	29176	14015	1642				2432		929	
2 – CCMN	832	2048	60	n/a	n/a	n/a	47	n/a	10	n/a
3 – CCL	547	1783	1				1		0	
4 – CPNN	2051		92				161			
5 – CPNM	850		34				19			
6 – CPL	122		5				11			
7 – FNN	1205	597	63				59		1	
8 – FMN	1216	292	68				64		0	
9 – FL	1115	847	50				69		9	
% missing	9.2	25.5	1.8				1.9		6.7	

CON : For UK, the subjective identity question of the OPCS census; for Germany, country of citizenship. Category 1 indicates number of missing cases.

EC9 : only derived if at least 6 of the 9 categories may be distinguished.

Table 4 : Ethnicity measures and associations with selected variables, ESS 2002

	IMM	LAN	VIS	MIN	CON	EC9	NONE
Sample: All ESS North-West European countries, n=21,782							
	<i>Bivariate categorical association statistics (eta*100):</i>						
Age in years	5*	8*	7*	8*		14*	
Years of educ	2*	4*	1	2*		8*	
ISEI	1	3*	1	1*		5*	
Use of internet	1	1	2*	4*		5*	
Left-right scale	5*	1	5*	7*		8*	
	<i>Regression model: ISEI = Gender + age + age-squared + years educ + [ethnicity]</i>						
R2	255	255	256	256		255	254
# sig dummies	-	-	-	-		- / -	
Influence oths.	none	none	none	none		none	
Sample: UK only, n=1958							
	<i>Bivariate categorical association statistics (eta*100):</i>						
Age in years	6*	10*	12*	15*	15*	20*	
Years of educ	16*	8*	15*	18*	18*	20*	
ISEI	4	3	3	2	2	7	
Use of internet	7*	7*	7*	11*	11*	13*	
Left-right scale	2	2	5*	4	4	7	
	<i>Regression model: ISEI = Gender + age + age-squared + years educ + [ethnicity]</i>						
R2	248	246	247	255	248	251	247
# sig dummies	-	-	-	-		- / - / -	
Influence oths.	none	none	none	none	none	none	
Sample: Germany only, n=2555							
	<i>Bivariate categorical association statistics (eta*100):</i>						
Age in years	7*	8*	9*	9*	10*	13*	
Years of educ	9*	10*	8*	4	11*	13*	
ISEI	7*	9*	6*	3	7*	10*	
Use of internet							
Left-right scale	2	3	4*	5*	5	8*	
	<i>Regression model: ISEI = Gender + age + age-squared + years educ + [ethnicity]</i>						
R2*100	344	345	344	343	343	344	344
# sig dummies		-					
Influence oths.	none	none	none	none	none	none	
<p>*: coefficient significant beyond 95% criteria. R2 is 'adjusted R-squared' from an OLS multiple regression in SPSS v12. # sig dummies: sign (& number) of effect(s) significant at 95%, for ethnic category dummy variable(s). 'Influence oths' indicates any of other predictors whose coefficients change from the model with no ethnicity information to the point of having non-overlapping 95% confidence intervals.</p>							

Table 5: SOR score derivations and model properties, ESS 2002

	North-West Europe sample, n=20899		UK only, n=1893		Germany, n=2503	
	M1 EC9	M2 EC9	M3 EC9	M4 CONU	M5 EC9	M6 CONG
SOR coeffs:						
Female	-0.06	0.10	0.04	0.03	-0.31	0.15
Age in years / 10	0.60*	0.21*	0.26*	0.12	0.58	0.50
Age squared / 1000	-1.11*	-0.48*	-0.46*	-0.21	-1.16*	-1.03*
Years of education / 10	0.10*	0.14*	0.03*	0.08*	-1.58*	-1.28*
Cohabiting	0.13	0.03	-0.15*	-0.04	-0.25	0.24
Currently working	-0.28*	-0.19*	-0.02	0.04	0.31	0.22
ISEI occ adv / 100	-0.93*	-0.28	0.12*	0.22	-1.30	-0.97
Self-employed	0.01	-0.08	0.08*	0.03	0.81*	0.36*
Female*cohabit	-0.23	-0.22*	0.04	-0.08	-0.13	-0.38
Country (M2, contrast Austria):						
Belg / Switz	-0.51*	0.74*				
Germany / Denmark	-0.57*	-1.32*				
Finland / UK	-2.36*	-0.36*				
Ireland / Luxem	-1.19*	1.98*				
Netherl / Norway	-0.85*	-1.12*				
Sweden		-0.36*				
* = Estimate significant at 95% criteria.						
SOR scores for EC9	EC9	EC9	EC9	CONU	EC9	CONG
1. CCNN	-239	-530	-253		-332	
2. CCMN	-220	-144	-152		-369	
3. CCL	-380†	-368†	-	-571	293†	-504
4. CPNN	-118	-22	-226	309†	-328	826†
5. CPMN	233	192	-143	643†	461†	-235
6. CPL	634	604	717†	22	324†	-87
7. FNN	-330	-203	484	-404	-387	
8. FMN	5	171	-241		33	
9. FL	414	300	-186		305	
† : Less than 20 cases in the SOR category CONU: 1=White; 2=Black-Caribbean; 3=Black-African; 4=Asian; 5=Other. CONG: 1=German citizen; 2=Turkish citizen; 3=Eastern European citizen; 4=Other citizen.						
SOR score effects			<i>Pearson's correlation*100</i>			
Age in years	-13*	-11*	-3	-10*	-12*	-10*
Years of educ	-2*	1	7*	14*	-10*	-10*
ISEI	-2*	1	-1	5*	-8*	-7*
Use of internet	2*	3*	3	8*	n\a	n\a
Left-right scale	-6*	-8*	-3	-3	-1	0
<i>Regression model: ISEI = Gender + age + age-squared + years educ + [ethnicity]</i>						
R2*100	256	256	253	253	345	343
Sig of SOR effect	0	0	-	0	0	0
Influence oths.	none	none	none	none	none	none
For notes on regression model summaries, see Table 4.						

References

- Ahmad, W. (1999). "Ethnic Statistics : Better than nothing or worse than nothing?" Pp. 124-131 in *Statistics in Society: The Arithmetic of Politics*. edited by Dorling, D. and Simpson, S. London: Arnold.
- Alba, R. and V. Nee. (2003). *Remaking the American Mainstream: Assimilation and Contemporary Immigration*. Harvard: Harvard University Press
- Allen, S. and M. Macey. (1990). Race and Ethnicity in the European Context. *British Journal of Sociology*. 41[3], 375-393.
- Anderson, J. A. (1984). "Regression and Ordered Categorical Variables," *Journal of the Royal Statistical Society Series B* 46(1):1-30.
- Aspinall, P. J. (2002). "Collective terminology to describe the minority ethnic population: The persistence of confusion and ambiguity in usage". *Sociology*. 36[4], 803-816.
- Aspinall, P. J. (2003). "The conceptualisation and categorisation of mixed race/ethnicity in Britain and North America: Identity options and the role of the state". *International Journal of Intercultural Relations*. 27[3], 269-296.
- Back, L. and J. Solomos. (1993). "Doing Research, Writing Politics: The dilemmas of political intervention in research on racism". *Economy and Society*. 22[2], 178-199.
- Ballard, R. (1997). "The construction of a conceptual vision : Ethnic groups and the 1991 UK census." *Ethnic and Racial Studies* 20:182.
- Banton, M. (1997). *Ethnic and Racial Consciousness*. London: Longman.
- Blackaby, D. H., S. Drinkwater, D. Leslie, and N. O'Leary. (1998). "Britain's Ethnic Communities." Pp. 38-62 In *An Investigation of Racial Disadvantage*, edited by Leslie, D., D. H. Blackaby, K. Clark, S. Drinkwater, P. Murphy, and N. O'Leary. Manchester: Manchester University Press.
- Bonifazi, C. and S. Strozza, (eds). (2003). "Integration of migrants in Europe: data sources and measurement in old and new receiving countries", special issue of *International Journal of Migration Studies* 40(152).
- Braun, M. and R. Uher. (2003). "The ISSP and its Approach to Background Variables." Pp. 33-48 In *Advances in Cross-National Comparison*, edited by Hoffmeyer-Zlotnick, J. H. P. and C. Wolf. New York: Kluwer.
- Brown, C. and J. Ritchie. (1981). *Focussed Enumeration : The development of a method for sampling ethnic minority groups*. London: Policy Studies Institute.

- Brown, M.S. (2000). "Quantifying the Muslim population in Europe: conceptual and data issues." *International Journal of Social Research Methodology* 3:87-102
- Chiswick, B. R. and P. W. Miller. (1995). "The Endogeneity between Language and Earnings - International Analyses." *Journal of Labor Economics* 13:246-288.
- de Vaus, D. (2002). *Surveys in Social Research, 5th Edition*. London: Routledge
- Evans, G. and A. Need. (2002). "Explaining ethnic polarization over attitudes towards minority rights in Eastern Europe: A multilevel analysis." *Social Science Research* 31:653-680.
- Favell, A. (2001). *Philosophies of Integration: Immigration and the Idea of Citizenship in France, 2nd Edition*. Palgrave: Macmillan
- Favell, A. (2003). Integration Nations: The Nation-State and Research on Immigrants in Western Europe. *Comparative Social Research*. 22, 13-42.
- Fenton, S. (1996). "Counting Ethnicity: Social Groups and Official Categories." Pp. 143-165 in *Interpreting Official Statistics*, edited by Levitas, R. and Guy W. London: Routledge.
- Ganzeboom, H. B. G. and D. J. Treiman. (1996). "Internationally Comparable Measures of Occupational Status for the 1988 International Standard Classification of Occupations." *Social Sciences Research* 25:201-235.
- Goldthorpe, J. H. (2000). *On Sociology : Numbers, narratives, and the integration of research and theory*. Oxford: Oxford University Press.
- Gunaratnam, Y. (2003). *Researching 'Race' and Ethnicity*. London: Sage.
- Green, A. E. and D. Owen. (1995). "Ethnic minority groups in regional and local labour markets in Britain : A review of data sources and associated issues." *Regional Studies* 29:729-735
- Hantrais, L. and S. Mangen. (1996). *Cross-National Research Methods in the Social Sciences*. London: Pinter
- Harkness, J., F. J. R. van de Vijver, and P. Ph. Mohler. (2003). *Cross-Cultural Survey Methods*. New York: Wiley
- Heath, A. F., D. McMahon, and J. Roberts. (2000). "Ethnic differences in the labour market: A comparison of the Samples of Anonymized Records and Labour Force Survey." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 163:303-339.
- Heckmann, F., R. D. Penn, and D. Schnapper. (2001). *Effectiveness of National Integration Strategies Towards Second Generation Migrant Youth in a Comparative Perspective - EFFNATIS*. Bamberg: European Forum for Migration Studies, University of Bamberg.

Hendrickx, J. (2000). "Special Restrictions in multinomial logistic regression," *Stata Technical Bulletin* STB-56:18-26.

Hoffmeyer-Zlotnik, J. H. P. (2003). "The Classification of Education as a Sociological Background Characteristic." Pp. 245-256 In *Advances in Cross-National Comparison: A European Working Book for Demographic and Socio-Economic Variables*, edited by Hoffmeyer-Zlotnik, J. H. P. and C. Wolf. New York: Kluwer Academic.

Hughes, A. O., S. Fenton, and C. E. Hine. (1995). "Strategies for sampling black and ethnic minority populations." *Journal of Public Health Medicine* 17:187-192.

Inglehart, R. (2003). *World Values Surveys and European Values Surveys 1981-4, 1990-3, 1995-7 [Computer file]*. Ann Arbor, MI : Institute for Social Research [Producer]; Ann Arbor, MI : Inter-university Consortium for Political and Social Research [Distributor]. 26.7.02

Jacobs, D. and J. Tillie, (eds). (2004). "Social Capital and Political Integration of Migrants", special issue of *Journal of Ethnic and Migration Studies*, 30(3).

Jamshidian, M. (2004). "Strategies for the Analysis of Incomplete Data". In *Handbook of Data Analysis*, edited by Hardy, M. and D. Bryman. London: Sage.

Jowell, R. (2003). *European Social Survey 2002/2003: Technical Report*. London: Centre for Comparative Social Surveys, City University

Kiecolt, K. J. and L. E. Nathan. (1985). *Secondary Analysis of Survey Data*. Thousand Oaks, Ca., Sage.

Lambert, P.S. (2002). "Quantitative Representations of Ethnic Difference?". *Paper presented to the ISA RC28 Spring meeting, Nuffield College, Oxford, April 10-13th*. (<http://www.nuff.ox.ac.uk/rc28/Papers/lambert.pdf>).

Lambert, P.S. and R. D. Penn. (2001). *SOR models and Ethnicity data in LIS and LES : Country by Country Report*. Syracuse University, Syracuse, New York 13244-1020: Luxembourg Employment Study Paper No. 21, Maxwell School of Citizenship and Public Affairs. (<http://www.lisproject.org/wpapers.htm>)

Lloyd, C. (1995). "International Comparisons in the Field of Ethnic Relations." Pp. 31-46 In *Racism, Ethnicity and Politics in Contemporary Europe*, edited by Hargreaves, A.G. and J. Leaman. Aldershot: Edward Elgar

Lynn, P. (2003). Developing quality standards for cross-national survey research: five approaches. *International Journal of Social Research Methodology*. 6[4], 323-336.

Marsh, C. (1982). *The Survey Method : The contribution of surveys to sociological explanation*. London: Allen and Unwin.

- Merritt, R. L. and S. Rokkan. (1966). *Comparing Nations: The Use of Quantitative Data in Cross-National Research*. New Haven: Yale University Press.
- Model, S., G. Fisher, and R. Silberman. (1999). "Black Caribbeans in comparative perspective." *Journal of Ethnic and Migration Studies* 25:187-212.
- Modood, T. (1991). "The Indian economic success - A challenge to some race-relations assumptions." *Policy and Politics* 19:177-189
- Modood, T., R. Berthoud, J. Lakey, J. Y. Nazroo, Patten Smith, S. Virdee, and S. Beishon. (1997). *Ethnic Minorities in Britain : Diversity and Disadvantage*. London: Policy Studies Institute.
- Modood, T., R. Berthoud, and J. Nazroo. (2002). "'Race', racism and ethnicity: A response to Ken Smith". *Sociology*. 36[2], 419-427.
- Owen, D. (1996). *Towards 2001 : Ethnic Minorities and the Census*. Warwick: Centre for Research in Ethnic Relations, University of Warwick.
- Panayi, P. (1999). *Outsiders : A History of European Minorities*. London: The Hambledon Press.
- Portes, A. and R. G. Rumbaut. (2001). *Legacies: The Story of the Immigrant Second Generation*. Berkeley and Los Angeles: University of California Press
- Prandy, K. (1979). "Ethnic discrimination in employment and housing," *Ethnic and Racial Studies* 2(1): 66-79.
- Punch, K. F. (2003). *Survey Research: The Basics*. London, Sage
- Rea, A., J. Wrench, and N. Ouali. (1999). "Discrimination and Diversity." In *Migrants, ethnic minorities and the labour market : Integration and Exclusion in Europe*, edited by Wrench, J., A. Rea, and N. Ouali. London: Palgrave.
- Sillitoe, K. and P. H. White. (1992). "Ethnic Group and the British Census : the search for a question." *Journal of the Royal Statistical Society, Series A : Statistics in Society* 155:141-163.
- Simpson, S. (1996). "Non-response to the 1991 census : The effect on ethnic group enumeration." In *Ethnicity in the 1991 Census : Volume 1*, edited by Coleman, D. and J. Salt. London: HMSO.
- Smith, D.M. and M. Blanc. (1995). "Some Comparative Aspects of Ethnicity and Citizenship in the European Union." Pp. 70-92 of *Migration, Citizenship and Ethno-National Identities in the European Union* edited by M. Martiniello. Aldershot: Avebury.
- Southworth, J. (1999). "The religious question: representing reality or compounding confusion?" Pp. 132-139 In *Statistics in Society: The Arithmetic of Politics*, edited by Dorling, D. and S. Simpson. London: Arnold.

- Steuer, M. (2003). *The Scientific Study of Society*. Boston: Kluwer Academic.
- Stille, F. (1999). "Ethnic Minorities and Immigrant Groups on the Labour Market : Introduction." *SYSDEM : Employment Observatory Trends* 32.
- SYSDEM. (2003). "Dealing with the Challenge of Immigration". *European Employment Observatory Review*. Autumn, 27-112.
- van Deth, J. W. (2003). "Using Published Survey Data." Pp. 329-346 In *Cross-Cultural Survey Methods*, edited by Harkness, J. A., F. J R. Van de Vijver, and P. Ph. Mohler. New York: Wiley.
- Van Tubergen, F., I. Maas, and H. Flap. (2004). "The Economic Incorporation of Immigrants in 18 Western Societies: Origin, Destination and Community Effects." (forthcoming in *American Sociological Review*).
- Wrench, J. and J. Solomos. (1993). *Racism and migration in Western Europe*. Oxford: Berg.