Data management and standardisation

Data management and standardisation: A methodological comment on using results from the UK Research Assessment Exercise 2008

Paul S. Lambert University of Stirling
Vernon Gayle

5th January 2009 [Edition 0.2]

DAMES Node, Technical Paper 2008-3

Technical Papers of the DAMES Node: Data Management through e-Social Science, http://www.dames.org.uk/publications.html. DAMES is an ESRC Research Node, Ref: RES-149-25-1066, based at the Universities of Stirling and Glasgow (the National e-Science Centre, www.nesc.ac.uk). The DAMES Node is a component of the ESRC National Centre for e-Social Science (www.ncess.ac.uk).

Version history:
Edition 0.1 published on 24th December 2008
Edition 0.2 published on 5th January 2009

Contents:

Abstract

This paper uses an analysis of UK RAE 2008 data in order to demonstrate methodological issues concerned with 'data management' and the analysis of standardised data. The paper is intended to be an accessible introduction for non-specialists to selected issues of data management, and the standardisation of variables, in quantitative social science research. The paper concludes with a brief analysis of RAE data which draws different conclusions to previously circulated results.

1. Introduction

1.1 Origins of this paper

In this paper we provide a brief analysis of results of the UK's 2008 Research Assessment Exercise (RAE). Our analysis is motivated by methodological issues relevant to activities of 'data management' and the use of standardised data in quantitative analyses. Readers who are only interested in the analytical results of our study may wish to go directly to section 3 which provides a short analytical review focusing upon the comparative rankings of Higher Educational Institutes (HEIs) and academic subject Units of Assessment (UOAs).

The results of the UK's 2008 RAE were published on 18th December 2008 (see RAE, 2008). In the immediate aftermath of their publication, many commentators observed the apparent confusion associated with the interpretation of the published scores. It was argued that the results represented complex, 'non-comparable' rankings, made inaccurate by institutional and subject oriented bias. Competing interpretations of the same data were easily identified, and there was considerable debate about the probable future impact of what were extremely expensive data to collect (e.g. Caulkin, 2008).

We argue that there are two principle explanations for the difficulties encountered in understanding the 2008 RAE results. The first concerns the non-standardised, categorical nature of the data (which is structured according to Higher Educational Institutes and by academic subject Units of Assessment). The second concerns the challenges of 'data management' in the analysis of complex categorical data which, we speculate, inhibited more effective analyses of the RAE data.

Both of these explanations concern methodological contributions within the remit of the Data Management through e-Social Science research Node based at the Universities of Stirling and Glasgow (DAMES, see www.dames.org.uk, is a research Node of the ESRC's National Centre for e-Social Science). The DAMES Node is concerned with providing resources and services concerned with the analysis of a number of specialist forms of social science data (including data concerned with occupations, educational qualifications, ethnicity, social care, and mental health). It also provides more generic resources concerned with social science data management.

Whilst the analysis of institutional level data on academic research is not central to the DAMES Node's substantive application areas, the UK's 2008 RAE nevertheless offers a informative – and widely understood – illustration of many of the relevant issues. Therefore, the following paper comprises a review of issues in handling the complex data generated through the RAE; a series of exemplar illustrations of data management routines which are instructive to this application; and a concluding analysis of the RAE results themselves which is intended to be accessible to a general audience. Further discussions and exemplar materials associated with this paper can be found on the DAMES Node website (www.dames.org.uk/rae2008/) .

1.2 Data management

Analysis of data such as the RAE results can be hindered by two challenges. One concerns the conceptual interpretation of categorical data and the identification of suitable statistical techniques for summarising it. A second concerns the practical challenges of drawing together complex data concerned with categorical units. Though prosaic, this latter challenge does in practice pose the bigger hurdle to research conduct in many academic fields. Difficulties include linking together different databases on related subjects, and finding a tractable organisation and summary for complex data resources. These tasks are typically called activities of 'data management', and are the central methodological focus of the DAMES research Node.

The UK 2008 RAE data is relatively small in quantity but it is still subject to significant data management requirements. Firstly, the data is organised in combinations of Higher Educational Institutes and Unit of Assessment subject groups within those HEIs. 2063 records are available, corresponding to data from 67 different Units of Assessment and 159 different HEI's[1]. Different analysts are interested in making different comparisons within and/or between different records. For instance, many users of the data are solely interested in a single record (that of their own host department), or in a limited range of comparisons (such as between all HEI's within their own UOA; or between all UOA's within their own HEI). Few users of the data are intrinsically interested in reviewing results across the full spectrum of the RAE. Accordingly, a typical method of publishing the RAE data is to present records in selected groups of UoA and/or HEI combinations (e.g. guardian.co.uk, 2008).

Secondly, other forms of external data may fruitfully inform analysis of the RAE records. For instance, aggregate information about research funding; publication volumes; student entry criteria; student ratings; HEI size; and faculty size are examples of data that is potentially available on the combinations of HEI-by-UOA units, and might be effectively linked with relevant records.

Effective analysis of the RAE data might involve extended 'data management' on the records by linking related data and restructuring the records. Nevertheless, preliminary reports on the RAE results were largely limited to basic statistical descriptions of the original datasets. In analyses below, we demonstrate examples of enhancing analysis of RAE data through data management activities.

[1] Most of the 159 HEI's have a moderate range of different UOA submissions, but none have submissions in all 67 UOAs. Across all HEI's, the mean number of different UOA's is 14.9 (standard deviation 13). The maximum number of submissions is 53 (University of Manchester). However the modal number of submissions is also the minimum number, 1 (covering 27 different institutes, mostly arts colleges and agricultural colleges). If (as we do in some analyses below) we exclude those HEI's with only one submission, the mean number of different UOA's per HEI is 18 (standard deviation 13). The number of different UOA's submitted by an institution does, in fact, correlate relatively strongly with the institution's RAE performance (large institutions achieving higher average quality profiles). See section 2.5 and [Stata-4] for further summaries of this data.

1.3 Standardisation and Categorical data

The quantitative analysis of categorical data has a long tradition of methodological reflection, especially in the social sciences, where most information is in some form categorical in nature. Categorical data can broadly be defined as information on the subjects of analysis which concerns a concept which is measured in a manner which relies upon distinctive boundaries between groups (categories). Since the boundaries between groups are inherently consequential, categorical data are often described as 'qualitative' in character, and are distinguished from measures which are 'quantitative' or 'metric' in nature (where numeric values indicate a subject's position within a finely graded dimension of difference)[2]. In the analysis of social survey data, for example, the subjects of analysis (survey respondents) typically have data stored on them which involves a mixture of categorical or qualitative measures (e.g. gender; region of residence; voting preference) and metric or quantitative measures (e.g. age in years; income).

The distinction between categorical and metric data is a staple of introductory texts on the analysis of quantitative data. Alternative statistical techniques are available to take account of data according to its categorical or metric properties. Of these, a traditional distinction between 'ordinal' and 'nominal' categorical data has significant implications for data analysis: ordinal categorical data incorporates information on a relative ranking of difference between categories (e.g. tax bands); nominal categorical data has no such apparent ranking (e.g. voting preference).

Despite the prevalence of categorical data, academic researchers in different disciplines are inconsistent in their approach to its statistical analysis. The implications of different 'levels of measurement' for arithmetic summaries were most famously highlighted by Stevens (1946), yet many contemporary analysts continue to ignore the issues, and present simplified arithmetic summaries of categorical records without further reflection (preliminary analyses of the RAE 2008 data arguably falls into this characterisation). Elsewhere, in some more numerate disciplines, highly specialised techniques for summarising certain types of categorical difference have been developed, but arguably at the cost that only relatively simple forms of categorical division are accommodated[3]. In other fields, it is commonplace to see categorical data subjected to a very limited range of statistical analyses (such as 'univariate' and 'bivariate' summaries), overtly because the multiple categories could not easily be incorporated into more complex multivariate analytical techniques.

The data provided by the UK RAE (available from RAE, 2008) are largely categorical in nature. Figure 1 gives an image of the basic data provided as downloaded in MS Excel format. The most significant variables (columns) are highlighted in Figure 1. The dataset contains 2363 rows of records. Each row indicates a unique combination

[2] Though this terminology should not be confused with the social science methodological tradition of the qualitative analysis of data

[3] Examples include the popular econometrics technique of 'propensity score matching' (e.g. Morgan & Winship, 2007) which is best suited to binary contrasts; and the popular sociological technique of event history analysis (e.g. Blossfeld et al., 2007) which is best suited to simple characterisations of an event's 'state space'.

from the 159 different Higher Educational Institutes and 67 different academic subject Units of Assessment. For each combination of HEI and UOA, the critical datum concerns information on the number of staff graded into one of five categories of academic research quality (4*, 3*, 2*, 1* and u/c). A small amount of additional data on each unit is also provided, namely the number of Full Time Equivalent staff contributing to the respective HEI-UOA ratings. For most analyses, therefore, there is one key set of outcome variables within the RAE data (the five indicators of quality profile, which are derived from categorical divisions between research quality ratings); and there are three important explanatory variables within the RAE data: measures of the HEI and UoA (both categorical in nature) and measures of the number of staff submitted for the relevant case (metric in nature).

Figure 1: Screenshot of the first cases from the MS Excel format version of the RAE data as downloaded from RAE (2008)

Microsoft Excel - main1.xls

File Edit View Insert Format Tools Data Window Help Adobe PDF

	A	B	C	D	E	F	G	H	I	J	K	L
1	2008 Research Assessment Exercise Results											
2								Overall quality profile				
3	Note: In this table, joint submissions are shown split by their constituent institutions.							(percentage of research activity at each quality level)				
4	Institution code	Institution name	Unit of assessment number	Unit of assessment name	Multiple submission letter	Multiple submission name	FTE Category A staff submitted	4*	3*	2*	1*	unclassified
5	H-0110	University of Birmingham	1	Cardiovascular Med			25.70	5	35	60	0	0
6	H-0112	University of Bristol	1	Cardiovascular Med			20.00	10	45	45	0	0
7	H-0114	University of Cambridge	1	Cardiovascular Med			14.00	35	50	10	5	0
8	H-0132	Imperial College London	1	Cardiovascular Med			64.27	20	45	30	5	0
9	H-0134	King's College London	1	Cardiovascular Med			32.40	20	60	20	0	0
10	H-0124	University of Leeds	1	Cardiovascular Med			30.00	5	40	55	0	0
11	H-0125	University of Leicester	1	Cardiovascular Med			21.00	10	45	40	5	0
12	H-0204	University of Manchester	1	Cardiovascular Med			24.20	15	60	25	0	0
13	H-0156	University of Oxford	1	Cardiovascular Med			18.60	45	40	15	0	0
14	H-0145	St George's Hospital Medical	1	Cardiovascular Med			16.86	5	45	40	5	5
15	H-0159	University of Sheffield	1	Cardiovascular Med			22.00	0	25	65	10	0
16	H-0168	University of Glasgow	1	Cardiovascular Med			50.60	15	50	30	5	0
17	H-0179	Cardiff University	1	Cardiovascular Med			19.40	5	35	40	20	0
18	H-0110	University of Birmingham	2	Cancer Studies			50.70	15	65	15	5	0
19	H-0114	University of Cambridge	2	Cancer Studies			33.50	35	45	15	5	0
20	H-0188	Institute of Cancer Research	2	Cancer Studies			65.51	35	40	25	0	0
21	H-0132	Imperial College London	2	Cancer Studies			51.55	15	60	20	5	0
22	H-0134	King's College London	2	Cancer Studies			30.00	5	50	45	0	0
23	H-0124	University of Leeds	2	Cancer Studies			48.70	15	65	20	0	0
24	H-0125	University of Leicester	2	Cancer Studies			27.05	5	50	40	0	5
25	H-0126	University of Liverpool	2	Cancer Studies			22.00	10	30	50	5	5
26	H-0204	University of Manchester	2	Cancer Studies			38.70	30	60	10	0	0
27	H-0154	University of Newcastle upor	2	Cancer Studies			32.93	15	75	10	0	0
28	H-0156	University of Oxford	2	Cancer Studies			38.75	25	50	20	5	0
29	H-0139	Queen Mary, University of Lo	2	Cancer Studies			33.30	15	70	15	0	0
30	H-0159	University of Sheffield	2	Cancer Studies			34.85	5	55	40	0	0
31	H-0160	University of Southampton	2	Cancer Studies			24.79	15	60	20	5	0
32	H-0149	University College London	2	Cancer Studies			40.40	25	50	15	5	5
33	H-0168	University of Glasgow	2	Cancer Studies			42.05	25	50	25	0	0
34	H-0179	Cardiff University	2	Cancer Studies			26.60	10	55	30	5	0
35	H-0184	Queen's University Belfast	2	Cancer Studies			37.00	10	40	45	5	0
36	H-0110	University of Birmingham	3	Infection and Immu			50.00	10	40	40	5	5
37	H-0112	University of Bristol	3	Infection and Immu			28.00	5	45	50	0	0
38	H-0114	University of Cambridge	3	Infection and Immu			46.00	35	45	15	0	5

Sheet1 simple

Ready Sum=376138.22

Most readers of the RAE data are interested in relative rankings of UOA-HEI combinations within and/or between different UOAs. For instance, a member of staff who was ennumerated within the Cancer Studies UOA group from, say, the University of Leeds, is likely to be most interested in the data from row 23 of Figure 1. They will also typically want to make comparisons between other records within their UOA (rows 18 to 35), and, perhaps, with other rows from the University of Leeds (such as row 10). However, such comparisons between rows are felt to be challenging for several reasons. First, within UOA's, the number of staff contributing to each ranking varies substantially between HEIs (see the variation in the 'FTE'

column of figure 1). Second, between UOA's, the criteria for making rating decisions may have varied. Lastly, the categorical nature of the RAE ranking groups means that conventional arithmetic summaries of the quality profile measures might not be appropriate. In practice, most early analyses of the RAE data have used a 'Grade Point Average' measure, which is the mean of the numeric values of the five RAE categories (assigned the values 4, 3, 2, 1 and 0 for 4*, 3*, 2*, 1* and u/c respectively)[4].

All of these concerns are, primarily, about what in quantitative research is usually termed 'standardisation'. Standardisation can broadly be described as the arithmetic techniques used for comparison of statistical summaries of different measures on a coordinated scale. In the case of the RAE results, there are three principle concerns with standardisation. One is over how to compare proportions in different RAE categories for a single record (i.e., should the Grade Point Average measure be used, or should some other arithmetic summary or summaries be used, perhaps giving more weight to certain categories such as the top 4* rating). Another is how to coherently compare summaries within and between UOAs in the context of different distributions in the number of FTE staff within UOA's. The last, and perhaps most challenging, is how to account for the potential differences in categorisation criteria employed across different UOA's[5]. Despite their obvious relevance, the majority of results hitherto published on the UK RAE performance have documented little or no attention to these topics of standardisation.

Such challenges are typical examples of problems in dealing with categorical data which are regularly encountered by academic social scientists (but are not consistently dealt with between applications). A component of the DAMES research Node is concerned with facilitating techniques for exploiting research data which are not consistently employed. To that end, we present below examples of analyses of the data associated with the RAE, describing in detail the approach to data management and the standardisation of categorical data we employ. Our analysis is, nevertheless, a brief, selective and superficial one. A full and extended analysis of the RAE data is a very substantial undertaking.

In summary, this paper and its analyses are presented as a means of illustrating the integration of data management and data analysis considerations which are central to progress in methodological research. In the text below we show examples of linking together external records with the RAE data, reorganising and restructuring the data, and employing suitable statistical techniques for the analysis of categorical data (see section 2). We then illustrate how these can contribute to a substantially improved analytical interpretation of the RAE results (see section 3).

[4] This type of numerical summary of categorical differences is, in fact, precisely what Steven's (1946) warned against.

[5] Many readers will recognise the obvious parallel with attempts to standardise exam and coursework grading criteria across different institutions and subjects (which is typically attempted by using systems of External Examiners and by publication of objective grading criteria). In a similar way, the RAE employed external objective criteria to demarcate the five RAE categories. In principle, these external criteria automatically ensure standardisation and comparability of RAE ratings across different UOAs. However, in practice, few would be persuaded that such comparability has been fully achieved (cf. Lipsset, 2008b). As with exam and coursework standardisation, there is ample empirical evidence of non-comparability between UOA grading criteria (for the RAE, see section 2 below).

2. A process log for the analysis of RAE 2008 data

2.1 Workflow of quantitative data analysis

The term workflow is increasingly used by research methodologists to describe the sequence of activities involved in a particular analytical project. In quantitative data analysis, interest has often turned to mechanisms for describing and organising workflows, such as in an influential recent publication by Scott Long (2008) which focuses upon command syntax in the Stata software package.

In the following sub-sections we present a simple workflow describing procedures involved in accessing RAE data; standardising the data by internal analysis and by linking it with other relevant resources; organising the data for efficient analysis; and undertaking data analysis tasks. A simplified summary of the results of these analyses is then presented in section 3.

Central to our workflow account is the careful preservation of a record (or 'log') of the tasks we undertake. This log is presented in summary form within the text below, and in more extended detail in our associated website www.dames.org.uk/rae2008. Since the primary tools of our analysis are the popular statistical packages SPSS and Stata, our principle method of logging our workflows is by recording the relevant SPSS and Stata command syntax used to achieve each step of the process. Such syntactical logs are overwhelmingly favoured by experienced analysts of quantitative data (e.g. Levesque & SPSS Inc, 2008; Scott Long, 2008; Kohler & Kreuter, 2009), and are critical to our approaches to supporting data management for social science research in the DAMES research Node[6]. Below, segments of SPSS and Stata syntax described in our account are highlighted in relevant tables and figures, using square brackets to indicate the relevant segment (for example, [SPSS-1] refers to the first segment of SPSS syntax we describe, and [Stata-2] the second segment of Stata syntax).

As this paper is motivated in large part by methodological intentions – to highlight approaches to data management and the standardisation of categorical data – in section 2 below we cover some simple approaches in these areas in relatively extended detail. Summary analyses are presented in section 3.

2.2 Accessing the RAE data

The RAE data in published online (RAE, 2008). The principle source of the full quantitative data is from a pdf format report[7] or in an MS Excel format data file (the latter is illustrated in Figure 1 above)[8]. The results are also available from online tables published within HEI and UoA groups (guardian.co.uk, 2008; RAE, 2008). These modes of distribution are typical of academic research data distributed online.

[6] In a related project we have published many introductory materials in analysing quantitative data using SPSS and Stata syntax, see http://www.longitudinal.stir.ac.uk/ .
[7] http://submissions.rae.ac.uk/results/outstore/RAEOutcomeFull.pdf
[8] http://submissions.rae.ac.uk/results/outstore/Main%20table%20of%202008%20RAE%20results.xls

The data has been released in formats which are anticipated to be easily reviewed or exploited for simplified statistical analysis. These formats do not however lend themselves readily to more extended statistical analysis without further analytical effort.

In order to produce the analyses presented below, we downloaded the MS Excel format data files, and transferred this information into the formats of the popular general purpose statistical packages SPSS and Stata. Table 2 shows images of the data involved, and the the SPSS syntax operations we used to achieve this. It illustrates opening the Excel format data and saving in SPSS and Stata formats [SPSS-1]. It also illustrates an exercise we undertook to reformat some of the variables from a text format record to numeric records (using the 'autorecode' facility in SPSS). The SPSS and Stata format files as produced below are available for download from our website www.dames.org.uk/rae2008/ .

Box 1: SPSS and Stata syntax: A note on file locations.

Throughout the examples below we use a procedure whereby we define an alias for the absolute 'path' locations to describe the folders in which relevant data files are stored on our own machine, thereby invoking suitable data files by using the alias rather than the full path. This aids transferability of the data files. Say we had saved our data in our folder 'c:\data\rae200\', then we would first define paths in SPSS and Stata as follows:

[SPSS]: `define !path1 () 'C:\data\rae2008\' !enddefine.`

[Stata]: `global path1 "C:\data\rae2008\"`

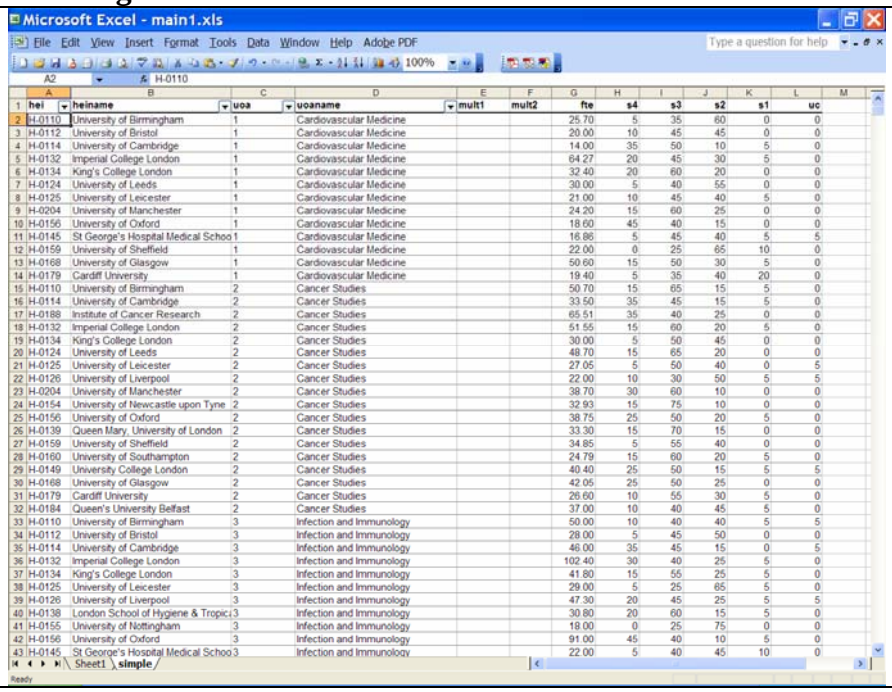
(The text 'path1' is arbitrary can could be any suitable alias). Subsequently, the relevant data files can be called in later routines with commands such as:

[SPSS]: `get file=!path1+"rae2008_1.sav"`

[Stata]: `use !path1+"rae2008_1.dta", clear`

Table 1: Reformatting the Excel data into SPSS and Stata formats

Panel 1: Original MS Excel data is reformatted removing header information



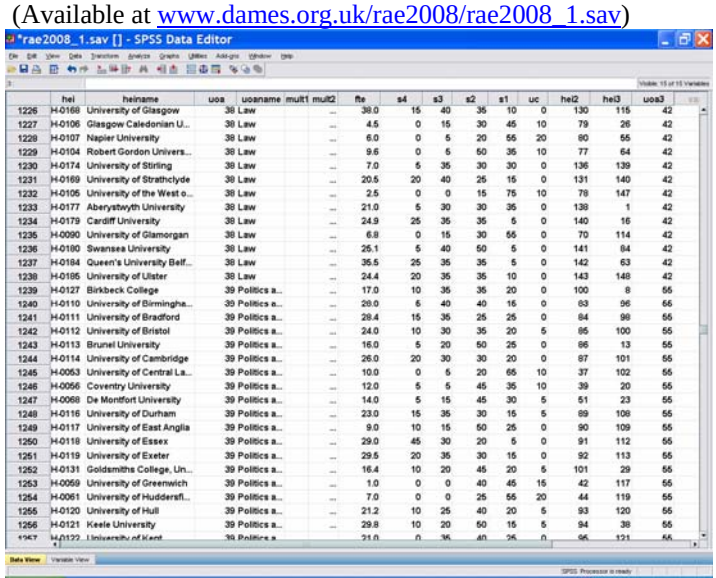
hei	heiname	uoa	uoaname	mult1	mult2	fte	s4	s3	s2	s1	uc
H-0110	University of Birmingham	1	Cardiovascular Medicine			25.70	5	35	60	0	0
H-0112	University of Bristol	1	Cardiovascular Medicine			20.00	10	45	45	0	0
H-0114	University of Cambridge	1	Cardiovascular Medicine			14.00	35	50	10	5	0
H-0132	Imperial College London	1	Cardiovascular Medicine			64.27	20	45	30	5	0
H-0134	King's College London	1	Cardiovascular Medicine			32.40	20	60	20	0	0
H-0124	University of Leeds	1	Cardiovascular Medicine			30.00	5	40	55	0	0
H-0125	University of Leicester	1	Cardiovascular Medicine			21.00	10	45	40	5	0
H-0204	University of Manchester	1	Cardiovascular Medicine			24.20	15	60	25	0	0
H-0156	University of Oxford	1	Cardiovascular Medicine			18.60	45	40	15	0	0
H-0145	St George's Hospital Medical Schoo	1	Cardiovascular Medicine			16.86	5	45	40	5	5
H-0159	University of Sheffield	1	Cardiovascular Medicine			22.00	0	25	65	10	0
H-0168	University of Glasgow	1	Cardiovascular Medicine			50.60	15	50	30	5	0
H-0179	Cardiff University	1	Cardiovascular Medicine			19.40	5	35	40	20	0
H-0110	University of Birmingham	2	Cancer Studies			50.70	15	65	15	5	0
H-0114	University of Cambridge	2	Cancer Studies			33.50	35	45	15	5	0
H-0188	Institute of Cancer Research	2	Cancer Studies			65.51	35	40	25	0	0
H-0132	Imperial College London	2	Cancer Studies			51.55	15	60	20	5	0
H-0134	King's College London	2	Cancer Studies			30.00	5	50	45	0	0
H-0124	University of Leeds	2	Cancer Studies			48.70	15	65	20	0	0
H-0125	University of Leicester	2	Cancer Studies			27.05	5	50	40	0	5
H-0126	University of Liverpool	2	Cancer Studies			22.00	10	30	50	5	5
H-0204	University of Manchester	2	Cancer Studies			38.70	30	60	10	0	0
H-0154	University of Newcastle upon Tyne	2	Cancer Studies			32.93	15	75	10	0	0
H-0156	University of Oxford	2	Cancer Studies			38.75	25	50	20	5	0
H-0139	Queen Mary, University of London	2	Cancer Studies			33.30	15	70	15	0	0
H-0159	University of Sheffield	2	Cancer Studies			34.85	5	55	40	0	0
H-0160	University of Southampton	2	Cancer Studies			24.79	15	60	20	5	0
H-0149	University College London	2	Cancer Studies			40.40	25	50	15	5	5
H-0168	University of Glasgow	2	Cancer Studies			42.05	25	50	25	0	0
H-0179	Cardiff University	2	Cancer Studies			26.60	10	55	30	5	0
H-0184	Queen's University Belfast	2	Cancer Studies			37.00	10	40	45	5	0
H-0110	University of Birmingham	3	Infection and Immunology			50.00	10	40	40	5	5
H-0112	University of Bristol	3	Infection and Immunology			28.00	5	45	50	0	0
H-0114	University of Cambridge	3	Infection and Immunology			46.00	35	45	15	0	5
H-0132	Imperial College London	3	Infection and Immunology			102.40	30	40	25	5	0
H-0134	King's College London	3	Infection and Immunology			41.80	15	55	25	5	0
H-0125	University of Leicester	3	Infection and Immunology			29.00	5	25	65	5	0
H-0126	University of Liverpool	3	Infection and Immunology			47.30	20	45	25	5	5
H-0138	London School of Hygiene & Tropic	3	Infection and Immunology			30.80	20	60	15	5	0
H-0155	University of Nottingham	3	Infection and Immunology			18.00	0	25	75	0	0
H-0156	University of Oxford	3	Infection and Immunology			91.00	45	40	10	5	0
H-0145	St George's Hospital Medical Schoo	3	Infection and Immunology			22.00	5	40	45	10	0

Panel 2: [SPSS-1]

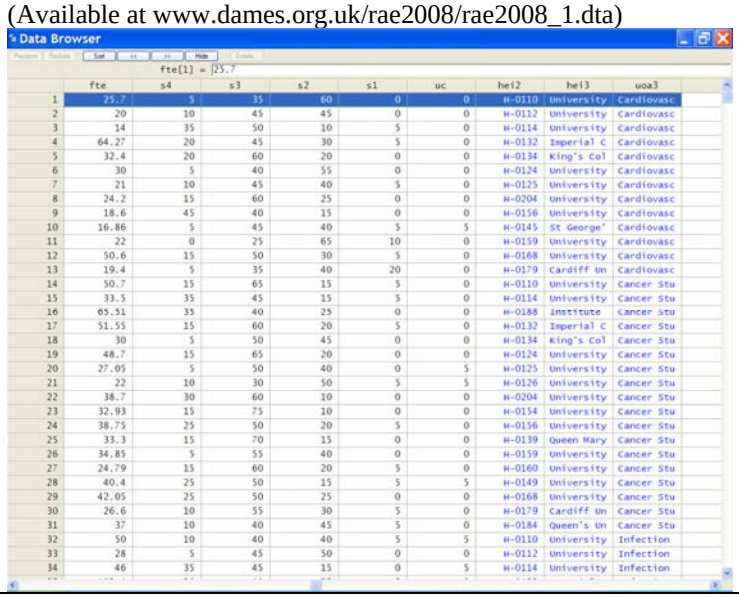
```
**************************************************.
*** Data access; conversion of string variables to numeric format;
**      export to SPSS and Stata data .
**************************************************.
get data /type=xls   /file=!path1+"main1.xls"
   /sheet=name "simple"
   /cellrange=range 'a1:l2364'  /readnames=on  .
descriptives var=all.
*.
autorecode var=hei heiname uoaname /into=hei2 hei3 uoa3 .
descriptives var=hei2 hei3 uoa uoa3 fte .
sav out=!path1+"rae2008_1.sav".
sav translate out=!path1+"rae2008_1.dta"
   /type=stata /version=8 /replace.
************************************************************.
```

Panel 3: Image of part of the SPSS data

(Available at www.dames.org.uk/rae2008/rae2008_1.sav)



	hei	heiname	uoa	uoaname	mult1	mult2	fte	s4	s3	s2	s1	uc	hei2	hei3	uoa3
1226	H-0168	University of Glasgow	38	Law		...	38.0	15	40	35	10	0	130	115	42
1227	H-0106	Glasgow Caledonian U...	38	Law		...	4.5	0	15	30	45	10	79	26	42
1228	H-0107	Napier University	38	Law		...	6.0	0	5	20	55	20	80	55	42
1229	H-0104	Robert Gordon Univers...	38	Law		...	9.6	0	5	50	35	10	77	64	42
1230	H-0174	University of Stirling	38	Law		...	7.0	5	35	30	30	0	136	139	42
1231	H-0169	University of Strathclyde	38	Law		...	20.5	20	40	25	15	0	131	140	42
1232	H-0105	University of the West o...	38	Law		...	2.5	0	0	15	75	10	78	147	42
1233	H-0177	Aberystwyth University	38	Law		...	21.0	5	30	30	35	0	138	1	42
1234	H-0179	Cardiff University	38	Law		...	24.9	25	35	35	5	0	140	16	42
1235	H-0090	University of Glamorgan	38	Law		...	6.8	0	15	30	55	0	70	114	42
1236	H-0180	Swansea University	38	Law		...	25.1	5	40	50	5	0	141	84	42
1237	H-0184	Queen's University Belf...	38	Law		...	35.5	25	35	35	5	0	142	63	42
1238	H-0185	University of Ulster	38	Law		...	24.4	20	35	35	10	0	143	148	42
1239	H-0127	Birkbeck College	39	Politics a...		...	17.0	10	35	35	20	0	100	8	55
1240	H-0110	University of Birmingha...	39	Politics a...		...	29.0	5	40	40	15	0	83	96	55
1241	H-0111	University of Bradford	39	Politics a...		...	29.4	15	35	25	25	0	84	98	55
1242	H-0112	University of Bristol	39	Politics a...		...	24.0	10	30	35	20	5	85	100	55
1243	H-0113	Brunel University	39	Politics a...		...	16.0	5	20	50	25	0	86	13	55
1244	H-0114	University of Cambridge	39	Politics a...		...	26.0	20	30	30	20	0	87	101	55
1245	H-0053	University of Central La...	39	Politics a...		...	10.0	0	5	20	65	10	37	102	55
1246	H-0056	Coventry University	39	Politics a...		...	12.0	5	5	45	35	10	39	20	55
1247	H-0068	De Montfort University	39	Politics a...		...	14.0	5	15	45	30	5	51	23	55
1248	H-0116	University of Durham	39	Politics a...		...	23.0	15	35	30	15	5	89	108	55
1249	H-0117	University of East Anglia	39	Politics a...		...	9.0	10	15	50	25	0	90	109	55
1250	H-0118	University of Essex	39	Politics a...		...	29.0	45	30	20	5	0	91	112	55
1251	H-0119	University of Exeter	39	Politics a...		...	29.5	20	35	30	15	0	92	113	55
1252	H-0131	Goldsmiths College, Un...	39	Politics a...		...	16.4	10	20	45	20	5	101	29	55
1253	H-0059	University of Greenwich	39	Politics a...		...	1.0	0	0	40	45	15	42	117	55
1254	H-0061	University of Huddersfi...	39	Politics a...		...	7.0	0	0	25	55	20	44	119	55
1255	H-0120	University of Hull	39	Politics a...		...	21.2	10	25	40	20	5	93	120	55
1256	H-0121	Keele University	39	Politics a...		...	29.8	10	20	50	15	5	94	38	55
1257	H-0122	University of Kent	39	Politics a...		...	21.0	0	35	40	25	0	95	121	55

Panel 4: Image of part of the Stata data	(Available at www.dames.org.uk/rae2008/rae2008_1.dta) 

Data Browser

fte[1] = 25.7

	fte	s4	s3	s2	s1	uc	hei2	hei3	uoa3
1	25.7	5	35	60	0	0	H-0110	University	Cardiovasc
2	20	10	45	45	0	0	H-0112	University	Cardiovasc
3	14	35	50	10	5	0	H-0114	University	Cardiovasc
4	64.27	20	45	30	5	0	H-0132	Imperial C	Cardiovasc
5	32.4	20	60	20	0	0	H-0134	King's Col	Cardiovasc
6	30	5	40	55	0	0	H-0124	University	Cardiovasc
7	21	10	45	40	5	0	H-0125	University	Cardiovasc
8	24.2	15	60	25	0	0	H-0204	University	Cardiovasc
9	18.6	45	40	15	0	0	H-0156	University	Cardiovasc
10	16.86	5	45	40	5	5	H-0145	St George'	Cardiovasc
11	22	0	25	65	10	0	H-0159	University	Cardiovasc
12	50.6	15	50	30	5	0	H-0168	University	Cardiovasc
13	19.4	5	35	40	20	0	H-0179	Cardiff Un	Cardiovasc
14	50.7	15	65	15	5	0	H-0110	University	Cancer Stu
15	33.5	35	45	15	5	0	H-0114	University	Cancer Stu
16	65.51	35	40	25	0	0	H-0188	Institute	Cancer Stu
17	51.55	15	60	20	5	0	H-0132	Imperial C	Cancer Stu
18	30	5	50	45	0	0	H-0134	King's Col	Cancer Stu
19	48.7	15	65	20	0	0	H-0124	University	Cancer Stu
20	27.05	5	50	40	0	5	H-0125	University	Cancer Stu
21	22	10	30	50	5	5	H-0126	University	Cancer Stu
22	38.7	30	60	10	0	0	H-0204	University	Cancer Stu
23	32.93	15	75	10	0	0	H-0154	University	Cancer Stu
24	38.75	25	50	20	5	0	H-0156	University	Cancer Stu
25	33.3	15	70	15	0	0	H-0139	Queen Mary	Cancer Stu
26	34.85	5	55	40	0	0	H-0159	University	Cancer Stu
27	24.79	15	60	20	5	0	H-0160	University	Cancer Stu
28	40.4	25	50	15	5	5	H-0149	University	Cancer Stu
29	42.05	25	50	25	0	0	H-0168	University	Cancer Stu
30	26.6	10	55	30	5	0	H-0179	Cardiff Un	Cancer Stu
31	37	10	40	45	5	0	H-0184	Queen's Un	Cancer Stu
32	50	10	40	40	5	5	H-0110	University	Infection
33	28	5	45	50	0	0	H-0112	University	Infection
34	46	35	45	15	0	5	H-0114	University	Infection

2.3 Exploring basic summary data

A first question concerns the basic distribution of values across RAE categories. The core data covers the percent of cases within each UOA/HEI record at each threshold. This data is commonly summarised by the 'GPA quality indicator', the arithmetic summary of the numeric values of the category indicators for each threshold. A further summary measure, an indicator of 'Research power', is also commonly used. It is calculated as the arithmetic product of the 'GPA quality indicator' and the number of FTE staff submitted by each UOA-HEI unit. Table 2 illustrates the calculation of these figures, and their presentation for certain selected subsets of the RAE data. It is these types of presentations that the large majority of initial analyses of RAE data have been restricted to.

Typically, previous descriptive analyses have stopped at the point of commentating on ranked positions within a selected group of records – in the records shown below, we present cases in order of 'GPA quality indicator' followed by 'Research power'. These are the indicators which have perhaps been most commonly employed across reports, though results have also been cited in terms of rank order of proportions of staff in 4*; or proportions of staff in 4* or 3*, categories. These and other variations are all legitimate alternative statistical presentations, and will in most instances tend to tell roughly the same story. Nevertheless, when simple alternative measures are juxtaposed, they have been widely interpreted as cynical statistical manipulations, and even downright lies (e.g. Lipsset, 2008b).

Table 2: RAE summary profiles

Panel 1: [Stata-2]

```
******** [Stata-2] Inspecting RAE quality profile measures
use $path1\rae2008_1.dta, clear
summarize
capture drop gpa_1
gen gpa_1 = (s4*4 + s3*3 + s2*2 + s1*1) / 100
capture drop power_1
gen power_1 = gpa_1*fte
summarize gpa_1 power
gsort -gpa_1 -power_1
tab uoaname if uoa==40
list heiname fte s* uc gpa_1 power_1 if uoa==40
numlabel _all, add
tab hei3
capture drop scot
gen scot=(hei3==25 | hei3==26 | hei3==27 | hei3==32 | ///
  hei3==55 | hei3==61 | hei3==64 | hei3==73 | ///
  hei3==86 | hei3==92 | hei3==93 | hei3==107 | hei3==115 | hei3==111 |  ///
        hei3==138 | hei3==139 | hei3==140 | hei3==147  )
tab heiname scot if scot==1
list heiname fte s* uc gpa_1 power_1 if uoa==40 & scot==1
list uoaname fte s* uc gpa_1 power_1 if hei3==139
************************************
```

Panel 2: subset of cases from UoA 40

This record shows responses for the highest ranked 35 records within UoA 40 (Social work and social policy and administration)

	heiname	fte	s4	s3	s2	s1	uc	gpa_1	power_1
21.	London School of Economics and Political Science	50.7	50	30	20	0	0	3.3	167.31
102.	University of Bath	23.7	35	40	25	0	0	3.1	73.47
180.	University of Southampton	31.5	35	35	25	5	0	3	94.5
221.	University of Kent	55.7	30	40	25	5	0	2.95	164.315
247.	University of Leeds	31.9	35	30	30	5	0	2.95	94.105
356.	University College London	7	15	60	25	0	0	2.9	20.3
374.	University of York	53.75	25	40	30	5	0	2.85	153.1875
382.	University of Edinburgh	43.9	30	35	25	10	0	2.85	125.115
445.	City University, London	3	20	45	35	0	0	2.85	8.549999
494.	University of Sheffield	26.6	20	45	30	5	0	2.8	74.48
502.	University of Oxford	22.83	20	50	20	10	0	2.8	63.924
551.	Lancaster University	42.25	20	40	35	5	0	2.75	116.1875
557.	Keele University	39.5	15	50	30	5	0	2.75	108.625
652.	University of Bristol	47.36	20	40	30	10	0	2.7	127.872
688.	University of Birmingham	26.12	15	45	35	5	0	2.7	70.524
709.	London South Bank University	18.8	15	45	35	5	0	2.7	50.76
730.	Nottingham Trent University	14.7	15	45	35	5	0	2.7	39.69
750.	University of Sussex	9	15	45	35	5	0	2.7	24.3
784.	University of Durham	35	15	45	30	10	0	2.65	92.75
811.	University of Warwick	22.8	10	50	35	5	0	2.65	60.42
819.	Queen's University Belfast	21.61	20	35	35	10	0	2.65	57.2665
840.	University of Ulster	15.6	10	50	35	5	0	2.65	41.34
853.	Open University	12.3	10	50	35	5	0	2.65	32.595
918.	University of Stirling	25.8	10	45	40	5	0	2.6	67.07999
951.	Swansea University	17.5	10	50	30	10	0	2.6	45.5
1034.	University of Plymouth	16.7	5	50	40	5	0	2.55	42.585
1036.	University of Bradford	16.4	10	40	45	5	0	2.55	41.82
1038.	University of East Anglia	16	10	45	35	10	0	2.55	40.8
1060.	University of Manchester	13	15	30	50	5	0	2.55	33.15
1124.	University of Nottingham	30.5	10	40	40	10	0	2.5	76.25
1181.	University of Bedfordshire	9.1	5	45	45	5	0	2.5	22.75
1186.	University of Reading	8.9	5	45	45	5	0	2.5	22.25
1213.	University of Salford	33.2	5	45	40	10	0	2.45	81.34
1241.	University of Glamorgan	18.1	5	40	50	5	0	2.45	44.345
1251.	Middlesex University	15.2	5	45	40	10	0	2.45	37.24

Panel 3: Subset of Scottish cases from UOA 40,

This record shows responses from panels within UoA 40 (Social work and social policy and administration), from Scottish HEI's only

	heiname	fte	s4	s3	s2	s1	scot	uc	gpa_1	power_1
382.	University of Edinburgh	43.9	30	35	25	10	1	0	2.85	125.115
918.	University of Stirling	25.8	10	45	40	5	1	0	2.6	67.07999
1497.	University of Strathclyde	10.15	5	35	45	15	1	0	2.3	23.345
1520.	University of Glasgow	4	5	35	45	15	1	0	2.3	9.2
1589.	University of Dundee	5.5	5	30	50	15	1	0	2.25	12.375
1847.	University of the West of Scotland	7	15	15	30	40	1	0	2.05	14.35

Panel 4: Subset of Cases from HEI = H-0174

This record shows all panels submitted by the University of Stirling

	uoaname	fte	s4	s3	s2	s1	scot	uc	gpa_1	power_1
285.	Philosophy	11	25	45	30	0	1	0	2.95	32.45
619.	Communication, Cultural and Media Studies	14	10	60	25	5	1	0	2.75	38.5
634.	Economics and Econometrics	7.6	15	45	40	0	1	0	2.75	20.9
918.	Social Work and Social Policy & Administration	25.8	10	45	40	5	1	0	2.6	67.07999
942.	Education	20.6	15	40	35	10	1	0	2.6	53.56
1033.	English Language and Literature	17	10	45	35	10	1	0	2.55	43.35
1145.	History	18.5	15	35	40	5	1	5	2.5	46.25
1150.	Nursing and Midwifery	16.6	20	30	35	10	1	5	2.5	41.5
1214.	Agriculture, Veterinary and Food Science	32.9	5	45	40	10	1	0	2.45	80.605
1355.	Sports-Related Studies	11.6	15	25	45	15	1	0	2.4	27.84
1362.	Computer Science and Informatics	10	5	40	45	10	1	0	2.4	24
1413.	Earth Systems and Environmental Sciences	19.8	5	35	50	10	1	0	2.35	46.53
1437.	Accounting and Finance	10.8	0	45	45	10	1	0	2.35	25.38
1464.	Business and Management Studies	31.9	10	30	40	20	1	0	2.3	73.37
1719.	Law	7	5	35	30	30	1	0	2.15	15.05
1833.	European Studies	12.03	5	25	45	20	1	5	2.05	24.6615
1969.	Psychology	24.2	5	10	55	30	1	0	1.9	45.98
2093.	Politics and International Studies	6	5	10	45	35	1	5	1.75	10.5

Key: fte = number of full time equivalent staff submitted for case; s4/s3/s2/s1/uc = proportion of submitted staff ranked within 4*, 3*, 2*, 1* and unclassified categories respectively; gpa_1 = 'Grade Point Average' (arithmetic mean of numeric category labels); power_1 = 'Research power' (gpa_1*fte)

2.4 Calculating averages and weighting by FTE

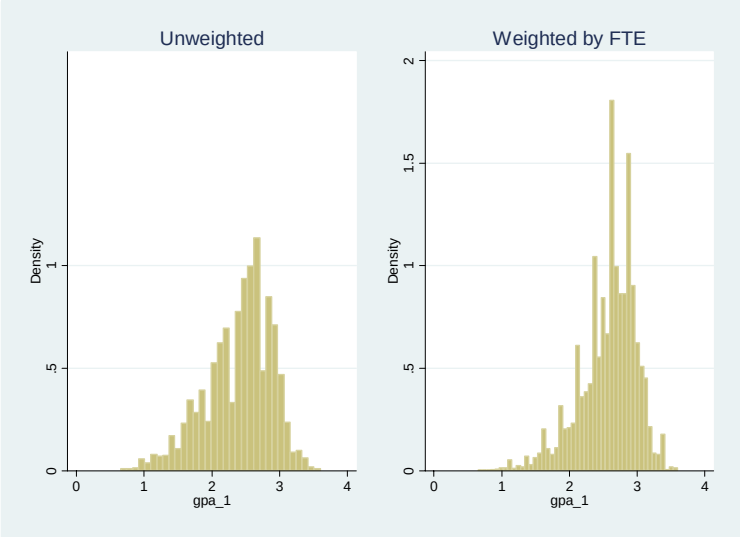
It is natural for many analyses to wish to present averages of the measures used above. For the next sections, we will concentrate only upon the commonly used measure of Grade Point Average (gpa_1). Across all or groups of the 2363 records, we could calculate the arithmetic mean, or other summary statistics, for this measure. Several reports of RAE grades have done this, though such averages would obviously be misleading given substantial differences in the number of FTE staff submitted to most panels. Therefore, more appropriate summary statistics may be calculated from weighted averages using the variable FTE as a population weight. Table 3 shows such a calculation for the gpa_1 measure[9]. In fact, the histograms in Table 3, panel 2, illustrate that unweighted average statistics are likely to over-represent units with lower Grade Point Average scores.

Table 3: Average measures of Grade Point Average scores

Panel 1: [Stata-3]

```
******** [Stata-3] Average measures of GPA
use $path1\rae2008_1.dta, clear
summarize
capture drop gpa_1
gen gpa_1 = (s4*4 + s3*3 + s2*2 + s1*1) / 100
histogram gpa_1, title(Unweighted) yscale( range(0 2))
graph save bit1, replace
gen ftei=floor(fte+1)
histogram gpa_1 [fweight=ftei], title(Weighted by FTE) yscale( range(0 2))
graph save bit2, replace
graph combine bit1.gph bit2.gph
graph export $path2\gpa_histograms_1.emf, as(emf) replace
```

[9] In this example, the shape of the distribution of GPA scores across records is such that the arithmetic mean can indeed be expected to give an informative summary of the overall distribution, with little difference, ordinarily, between the arithmetic mean and the median.

Panel 2	Histograms of the unweighted and weighted gpa_1 distributions 

2.5 Internal summaries

Many variables within the RAE data file could themselves be usefully summarised and linked 'internally' to enhance the results of analyses. This sort of data management operation is particularly easy to achieve, yet is often not undertaken. We show numerous examples of such manipulations in subsequent sections, but to begin with we show an illustrative example using data on the number of different UOA submissions entered by each University (also mentioned briefly in section 1.2). This simple information proves remarkably informative – it is generally the case that HEIs entering larger numbers of UOA records are likely to average higher GPA scores. This summary data also gives us a useful reminder that the overall RAE database contains some potentially 'outlying' cases, in that there are a large number of specialist institutions with only one UOA submission (mainly arts colleges and agricultural colleges) who might usefully be separated from some other analyses). Table 4 shows Stata syntax illustrating this summary process [Stata-4], as well as summary outputs which highlight the (possible) significance of this derived measure.

Table 4: Calculating average numbers of UOA submissions within HEIs

Panel 1: [Stata-4]

```
******** [Stata-4]  Mean numbers of UoA submissions by HEIs (ignores multiple submissions):
use $path1\rae2008_1.dta, clear
gen uoa_u=1
collapse (sum) uoa_u, by(hei3)
summarize
tab uoa_u
codebook uoa_u
tab hei3 if uoa_u == 1
tab hei3 if uoa_u == 53
codebook uoa_u if uoa_u >= 2
histogram uoa_u , title("Number of UoA's, all HEI's") yscale( range(0 0.08))
graph save bit1, replace
histogram uoa_u if uoa_u >= 2, title("Number of UoA's, exlcuding specialists") ///
                        yscale( range(0 0.08))
graph save bit2, replace
graph combine bit1.gph bit2.gph, title(RAE 2008)
graph export $path2\uoas_1.emf, as(emf) replace
table hei3 if uoa_u >= 40, c(mean uoa_u)
```

Panel 2

This panel shows all HEIs with only one UOA submission

hei3	Freq.	Percent	Cum.
Armagh Observatory	1	3.70	3.70
Arts Institute at Bournemouth	1	3.70	7.41
British Institute in Paris	1	3.70	11.11
Central School of Speech and Drama	1	3.70	14.81
Courtauld Institute of Art	1	3.70	18.52
Glasgow School of Art	1	3.70	22.22
Guildhall School of Music & Drama	1	3.70	25.93
Harper Adams University College	1	3.70	29.63
Institute of Education	1	3.70	33.33
Institute of Zoology	1	3.70	37.04
Leeds College of Music	1	3.70	40.74
London Business School	1	3.70	44.44
Norwich University College of the Arts	1	3.70	48.15
Rose Bruford College	1	3.70	51.85
Royal Academy of Music	1	3.70	55.56
Royal Agricultural College	1	3.70	59.26
Royal College of Art	1	3.70	62.96
Royal College of Music	1	3.70	66.67
Royal Northern College of Music	1	3.70	70.37
Royal Scottish Academy of Music and Dra	1	3.70	74.07
Royal Veterinary College	1	3.70	77.78
School of Pharmacy	1	3.70	81.48
Stranmillis University College	1	3.70	85.19
University College Falmouth	1	3.70	88.89
University for the Creative Arts	1	3.70	92.59
University Marine Biological Station, M	1	3.70	96.30
University of Wales Centre for Advanced	1	3.70	100.00
Total	27	100.00	

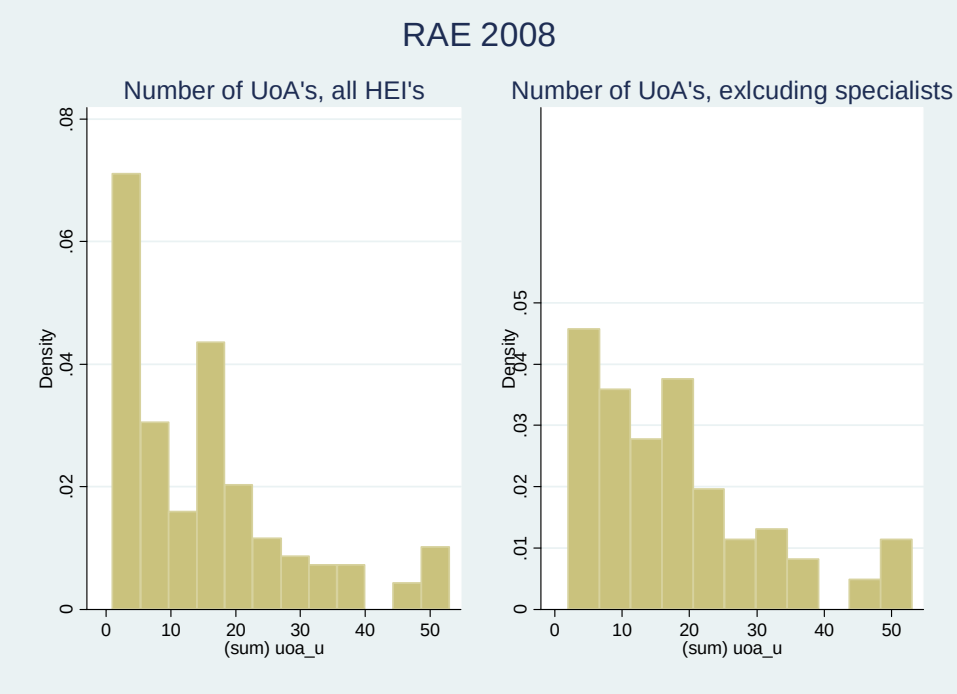
Panel 3

This panel shows the HEIs with over 40 different UOA submissions

hei3	mean(uoa_u)
University College London	49
University of Birmingham	49
University of Bristol	48
University of Cambridge	50
University of Glasgow	49
University of Leeds	46
University of Manchester	53
University of Nottingham	47
University of Oxford	50
University of Sheffield	49

Panel 4

This panel summarises the distribution of the number of UOAs across HEIs



2.6 Accessing and linking with external data

The RAE database contains at least two measures, the HEI identifier and the UOA identifier, which could be linked with external data about the relevant HEI and UOA groups. In fact, it is conceivable that all sorts of aggregate data at the HEI and/or the UOA level may illuminate our analysis of the breakdown of RAE gradings – a few interesting possibilities might be measures of university facilities; grant income; student entry levels; faculty demographics; staff salaries; RAE panel members' data. With structured data of the form described above, performing such 'deterministic' linkages is straightforward: information about HEI's and/or UoA's can be linked on the unique identifiers given for HEIs (the codes such as 'H-0174' for University of Stirling) or the identifiers given for the units of assessment. The Higher Education Statistics Agency (2008), for instance, is a useful source of many statistical summaries at this aggregate level.

For our own convenience, we curtail our use of external data linkage to a single example. We use a database that we have ourselves authored, and have posted for public access in Stata format at: http://www.dames.org.uk/rae2008/uni_typology.dta . This database gives a listing of HEI units, and data on them comprising records which indicate whether or not the HEI falls into one of a number of categories concerned with the era in which the university was made, and/or the type of affiliation the University holds with other institutions.

The Stata exert [Stata-5] show in Table 5 demonstrates this linkage, the end point being that we have a revised dataset comprising the original RAE data plus a number of derived measures (indicators of HEI type). We see from the weighted average gpa_1 distributions that the different HEI types have very different RAE performance patterns.

Table 5: Linking with an external dataset

Panel 1: [Stata-5]

```
******** [Stata-5]
** a) Data linkage and enhancements: RAE data linked with number of UoA's per
*   institution and external database of University typologies
use $path1\rae2008_1.dta, clear
summarize
gen uoa_u=1
collapse (sum) uoa_u, by(hei3)
summarize
sort hei3
sav $path9\m1.dta, replace
use $path1\uni_typology.dta, clear
summarize
sort hei3
sav $path9\m2.dta, replace
use $path1\rae2008_1.dta, clear
summarize
sort hei3
merge hei3 using $path9\m1.dta
drop _merge
summarize
sort hei3
merge hei3 using $path9\m2.dta
drop _merge
sav $path1\rae2008_2.dta, replace
```

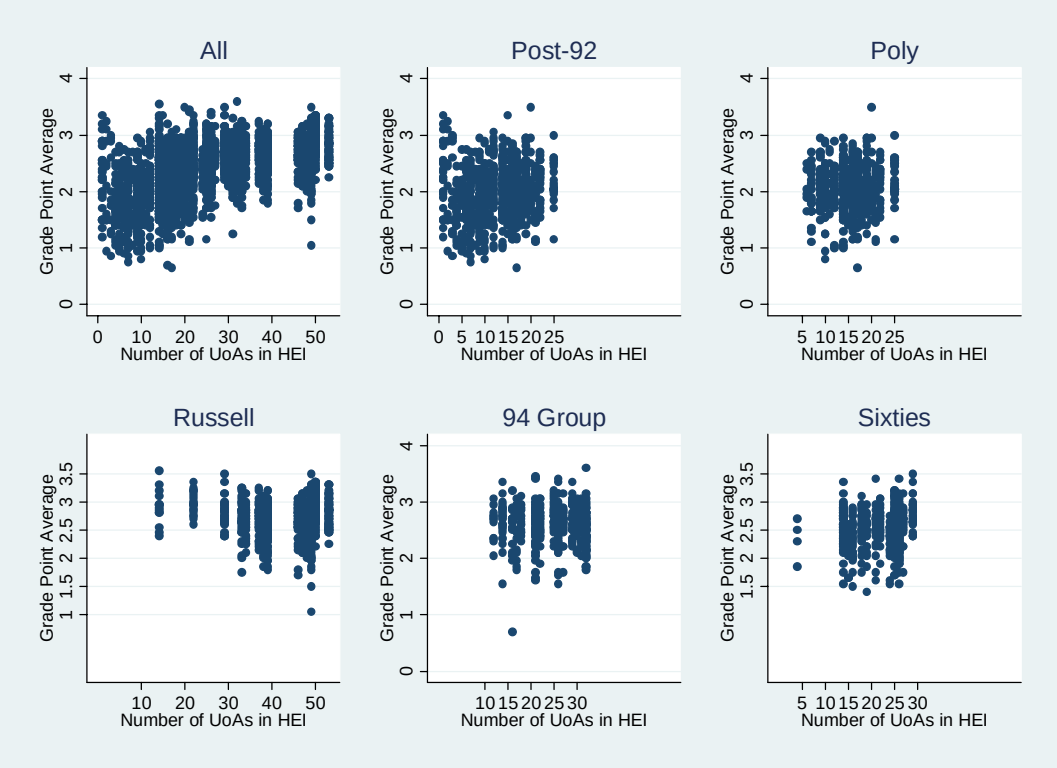
```
** b) Data used to show difference between university types
use $path1\rae2008_2.dta, replace
summarize
table oxbridge [aweight=fte], c(mean s4 mean s3 mean s2 mean s1)
table ancients [aweight=fte], c(mean s4 mean s3 mean s2 mean s1)
table civic [aweight=fte], c(mean s4 mean s3 mean s2 mean s1)
table redbrick [aweight=fte], c(mean s4 mean s3 mean s2 mean s1)
table plate [aweight=fte], c(mean s4 mean s3 mean s2 mean s1)
table sixties [aweight=fte], c(mean s4 mean s3 mean s2 mean s1)
table russell [aweight=fte], c(mean s4 mean s3 mean s2 mean s1)
table group94 [aweight=fte], c(mean s4 mean s3 mean s2 mean s1)
table pre92 [aweight=fte], c(mean s4 mean s3 mean s2 mean s1)
table poly [aweight=fte], c(mean s4 mean s3 mean s2 mean s1)
table nonpoly [aweight=fte], c(mean s4 mean s3 mean s2 mean s1)
table nonuni [aweight=fte], c(mean s4 mean s3 mean s2 mean s1)
capture drop wt
gen wt=s4*(fte/100)
graph hbar (mean) oxbridge ancients civic redbrick plate sixties russell ///
  group94 pre92 poly nonpoly ///
     [aweight=wt], title("Proportion of staff in 4-star category from..")

** c) Correlates between GPA, number of UoA submissions, & Unitypes
use $path1\rae2008_2.dta, clear
capture drop gpa_1
gen gpa_1 = (s4*4 + s3*3 + s2*2 + s1*1) / 100
summarize
label variable uoa_u "Number of UoAs in HEI"
label variable gpa_1 "Grade Point Average"
scatter gpa_1 uoa_u
correlate gpa_1 uoa_u oxbridge ancients russell
scatter gpa_1 uoa_u, title(All) yscale( range(0 4))  xscale( range(0 50))
graph save bit0 , replace
scatter gpa_1 uoa_u if russell==1, title(Russell) yscale( range(0 4))  xscale( range(0 50))
graph save bit1 , replace
scatter gpa_1 uoa_u if sixties==1, title(Sixties) yscale( range(0 4))  xscale( range(0 50))
graph save bit2 , replace
scatter gpa_1 uoa_u if group94==1, title(94 Group) yscale(range(0 4))  xscale( range(0 50))
graph save bit3 , replace
scatter gpa_1 uoa_u if poly==1, title(Poly) yscale( range(0 4))  xscale( range(0 50))
graph save bit4 , replace
scatter gpa_1 uoa_u if pre92==0, title(Post-92) yscale( range(0 4))  xscale( range(0 50))
graph save bit5 , replace
graph combine bit0.gph bit5.gph bit4.gph bit1.gph bit3.gph bit2.gph
graph export $path2\uoas_2.emf, as(emf) replace
************************************
```

Panel 2:

	Proportion of staff at given ratings by HEI type			
	4*	3*	2*	1*
Oxbridge	32	39	24	5
Ancients, Medieval, Renn.	26	40	27	7
Civic Universities	18	42	32	7
Red Brick	17	41	33	9
Plateglass	19	41	32	8
Royal charter in 1960's	16	38	34	11
Russell group	22	41	30	7
1994 Group	18	39	33	9
Pre-1992	19	40	32	8
Post-92 former polytechnic	7	27	40	23
Post-92 not former poly.	7	20	37	29
Non-standard institution	22	32	29	13

Panel 3: Relation between GPA score and Number of UoAs, split by selected HEI types.



2.7 Standardising categorical data across potentially different criteria

There are various circumstances in which researchers may wish to make comparisons between RAE results achieved in different UoA panels. Commonly, reviews within a particular HIE may wish to establish relatively more and less successful submissions with the institution. Hitherto, the majority of published comparisons have involved comparing the ratings of 'Grade Point Average' and 'Research Power' within an HEI (such as in the example of Table 2, panel 4, which shows all UoA results within the University of Stirling). However, as noted above, there are potential limitations in the consistency of criteria between UoAs.

There are two common standardisation approaches that may be taken to comparing between different UoAs. A first would be to calculate arithmetic standardisations within each UoA, in order that the mean Grade Point Average across UoAs is set to an equivalent value. This makes for easy comparison of a unit's relative attainment within its own UoA. The first panel of Table 6 presents Stata syntax to achieve this standardisation. It generates a measure ('gpa_2') with (arbitrarily) mean 50 and standard deviation 15 for the population of FTE staff submitted within each panel; accordingly, an individual unit achieving above or below 50 is on average achieving above or below the average for the panel. This comparison proves instructive for the

University of Stirling example, since at least one UoA group (Economics) achieves a relatively high Grade Point Average for the institution, but, when standardised within its UoA, performed relatively badly. Such standardisations within UoA categories are both instructive, and relatively easy to perform (see Table 6, panel 1); however they have largely been absent from discourses on the RAE hitherto.

Table 6: Grade Point Average standardisation within UoA's

Panel 1: [Stata-6]

```
******** [Stata-6]
** Mean standardisations using conventional gpa-measure
use $path1\rae2008_2.dta, clear
capture drop gpa_1
gen gpa_1 = (s4*4 + s3*3 + s2*2 + s1*1) / 100
sav $path9\m1.dta, replace
summarize gpa_1
collapse (mean) mean1=gpa_1 (sd) sd1=gpa_1 [aweight=fte], by(uoa3)
summarize
sort uoa3
sav $path9\m2.dta, replace
use $path9\m1.dta, clear
sort uoa3
merge uoa3 using $path9\m2.dta
tab _merge
drop _merge
capture drop gpa_2
gen gpa_2 = 50 + (15*((gpa_1 - mean1) / sd1))
summarize gpa_2 [aweight=fte]
* Example: standardised subject rankings at Stirling university
list uoa3 s4 s3 s2 s1 fte gpa* if hei3==139
table uoa3 if hei3==139, c(mean s4 mean s3 mean fte mean gpa_1 mean
gpa_2) format(%4.2f)
* Example: ranking within 1960 univs using within uoa zscore graph
hbar (mean) gpa_2 if sixties==1 [aweight=fte], ///
 over(hei3, label(labsize(vsmall)) sort(1) ) ///
    title(1960's: HIE mean gpa's standardised within UoAs)
```

Panel 2: GPA and within-UoA standardised GPA for University of Stirling

uoa3	mean(s4)	mean(s3)	mean(fte)	__000005	__000006
Accounting and Finance	0.00	45.00	10.80	2.35	50.28
Agriculture, Veterinary and Food Science	5.00	45.00	32.90	2.45	52.27
Business and Management Studies	10.00	30.00	31.90	2.30	41.17
Communication, Cultural and Media Studie	10.00	60.00	14.00	2.75	55.04
Computer Science and Informatics	5.00	40.00	10.00	2.40	37.50
Earth Systems and Environmental Sciences	5.00	35.00	19.80	2.35	30.83
Economics and Econometrics	15.00	45.00	7.60	2.75	37.97
Education	15.00	40.00	20.60	2.60	59.37
English Language and Literature	10.00	45.00	17.00	2.55	45.71
European Studies	5.00	25.00	12.03	2.05	44.67
History	15.00	35.00	18.50	2.50	41.85
Law	5.00	35.00	7.00	2.15	38.42
Nursing and Midwifery	20.00	30.00	16.60	2.50	50.13
Philosophy	25.00	45.00	11.00	2.95	59.96
Politics and International Studies	5.00	10.00	6.00	1.75	28.79
Psychology	5.00	10.00	24.20	1.90	34.69
Social Work and Social Policy & Administ	10.00	45.00	25.80	2.60	50.89
Sports-Related Studies	15.00	25.00	11.60	2.40	60.20

Key: As table 2, panel 4, plus: _00005 = conventional GPA average; _000006 = unit level gpa standardised within UoA average (to mean 50, sd 15).

Yet standardisation within the distribution of each UoA remains unsatisfactory, since it is widely anticipated that different UoA's operate with different average levels of research quality. Given this expectation, it may well be that a lower relative ranking within a UoA with higher standards is consistent with better research quality than is a higher relative ranking within a UoA with lower research standards. Of course, the

external criteria for RAE category rankings could in principle directly prevent this (i.e., UoA's with higher general research quality would simply have higher general GPA scores). Many users of RAE data may wish this property of equivalence to hold true, but it would take a considerable leap of faith to believe that the existing RAE rankings achieve such equivalence unproblematically.

A simple empirical demonstration of the likely problem is shown in Table 7. The weighted average RAE category scores and Grade Point Averages are shown for three Units of Assessment, Cardiovascular Medicine; Infection and Immunology; and Nursing and Midwifery. Submissions in the first two of these units are concentrated in HEIs with higher average RAE results generally, and from departments with high student intake criteria and high levels of staff credentials. Submissions in the latter unit are much more likely to come from institutions with lower overall RAE results, and from departments with relatively low student intake criteria, and lower staff credentials. Although a normative judgement, it might reasonably be expected that the research conducted in the former two panels would be, on average, substantially stronger than research in the latter group. As seen in Table 7, however, the RAE results do not bear this out strongly. Arithmetic differences between the overall UoA profiles are slight (especially when compared to the more substantial distributions between individual records). The profile of the Nursing and Midwifery UoA is slightly lower than the other two, but the difference is nothing like the order of magnitude we might expect. Of course, such an isolated and pejorative example doesn't prove the inconsistency of RAE rankings between Units of Assessment, but merely points in that direction. Many experienced academic researchers will be readily persuaded that the RAE rankings across different UoA's do not employ exactly equivalent criteria, though some readers will no doubt maintain the principle of equivalence in rankings for ideological reasons (e.g. Lipsset, 2008a), and only a much fuller evaluation of data on HEI performance indicators would be reasonable refutation.

Table 7: Illustration of FTE weighted RAE distributions across total UoAs

Panel 1: [Stata-7]

```
******** [Stata-7] : Comparison of three related panels
use $path1\rae2008_2.dta, clear
numlabel _all, add
capture drop gpa_1
gen gpa_1 = (s4*4 + s3*3 + s2*2 + s1*1) / 100
tab uoa3
table hei3 if uoa3==40
table hei3 if uoa3==14
table hei3 if uoa3==49
table uoa3 if uoa3==49 | uoa3==14 | uoa3==40 [aweight=fte] , ///
    c(mean s4 mean s3 mean s2 mean s1 mean gpa_1) format(%6.3g)
************************************
```

Panel 2: Indicative results comparing three UoA's

uoa3	mean(s4)	mean(s3)	mean(s2)	mean(s1)	mean(gpa_1)
14. Cardiovascular Medicine	14.6	45.1	36	4.02	2.7
40. Infection and Immunology	20.7	44.6	29.1	4.46	2.79
49. Nursing and Midwifery	19.7	32.7	29.7	13.2	2.5

If we are concerned about comparability across UoA profiles, how should we proceed further? The trite methodological point of this review is to argue that some fluency in data management approaches and methods of summarising categorical data offers a

major contribution. To whit, there is much information within the RAE data itself which could give us plausible measures of the relative research quality of a UoA as a whole, and there is even more external data on HEI's and UoA's which could be linked with the RAE database and readily deployed for comparative purposes. In Table 8, we show the calculation of 2 indexes which might be plausible indicators of a UoA's overall 'research standards'. Both measures rely on data at the HEI level, though other measures at UoA level could also have been employed here (and would probably be more persuasive). One measure is simply the proportion of staff submitted within the UoA working in a Russell group university (as seen in Table 5, panel 4, the Russell group Universities have consistently higher RAE evaluations on average). The second measure is an average of an average, namely the average HEI level 'Grade Point Average' held be the institution from which staff submitted within the UoA were hosted (i.e., the higher this average, the higher the overall RAE rankings of the institutions in which staff within the UoA typically worked).

Both of these institutionally based measures exploit a popular sociological approach to understanding stratification and inequality – the principles of social interaction distance. In many social arrangements of inequality, it can be show that networks of social interaction offer strong empirical indicators of the structure of stratification itself (e.g. Bottero, 2005). From this perspective, knowing about the average properties of the institution that surrounds an individual scholar tells us (on average) about the scholar's own research quality[10].

Table 8: Calculation of UoA level research quality indicators

Panel 1: [Stata-8]

```
***** [Stata-8]  Calculation of UoA level summary measures and mapping to unit level data
use $path1\rae2008_2.dta, clear
capture drop gpa_1
gen gpa_1 = (s4*4 + s3*3 + s2*2 + s1*1) / 100
sav $path9\m1.dta, replace
summarize gpa_1 russell
collapse (mean) mean1=gpa_1 (sd) sd1=gpa_1 (mean) rusmean=russell [aweight=fte], by(uoa3)
summarize
label variable rusmean "Proportion of UoA members from Russell group HEIs"
sort uoa3
sav $path9\m2.dta, replace
use $path9\m1.dta, clear
sort uoa3
merge uoa3 using $path9\m2.dta
tab _merge
drop _merge
capture drop gpa_2
gen gpa_2 = 50 + (15*((gpa_1 - mean1) / sd1))
label variable gpa_2 "Within UoA standardised GPA"
summarize gpa_2
summarize gpa*
summarize gpa* [aweight=fte]
sort uoa3
sav $path9\temp2.dta, replace
summarize gpa_2 fte
collapse (mean) gpa_2u=gpa_2 [aweight=fte], by(hei3)
summarize
sort hei3
```

[10] In fact, on a point of methodological interest, such institutional level structure can be discerned from the distribution of departments across HEIs alone (without inputting any additional data on RAE results). If a simple correspondence analysis is undertaken between HEI on the one hand, and UoA on the other, a dimension emerges which apparently represents 'research quality' and ranks both HEI's and UoA's within that dimension in a fashion which correlates highly with RAE based rankings.

```
label variable gpa_2u "HEI level average of within UoA standardised GPA"
sav $path9\m2.dta, replace
use $path9\temp2.dta, clear
sort hei3
merge hei3 using $path9\m2.dta
tab _merge
drop _merge
collapse (mean) unit_g=gpa_2u [aweight=fte], by(uoa3)
label variable unit_g "UOA level average of HEI level standardised GPAs"
summarize
sort uoa3
sav $path9\m3.dta, replace
use $path9\temp2.dta, clear
sort uoa3
merge uoa3 using $path9\m3.dta
tab _merge
drop _merge
sort hei3
merge hei3 using $path9\m2.dta
tab _merge
drop _merge
summarize
* Some illustrative results
table uoa3 [aweight=fte], c(mean gpa_1 mean gpa_2 mean unit_g sd unit_g)
table uoa3 [aweight=fte], c(mean gpa_1 mean unit_g mean rusmean mean russell)
correlate unit_g rusmean
graph hbar (mean) unit_g , over(uoa3, label(labsize(tiny)) sort(1) )
graph save bit1, replace
graph hbar (mean) rusmean , over(uoa3, label(labsize(tiny)) sort(1) )
graph save bit2, replace
graph hbar (mean) unit_g if sixties==1, over(hei3, label(labsize(tiny)) sort(1) )
graph save bit3, replace
graph hbar (mean) rusmean if sixties==1, over(hei3, label(labsize(tiny)) sort(1) )
graph save bit4, replace
graph combine bit1.gph bit2.gph, title(Ranking of UoAs by host HEI averages)
graph export $path2\uao_rank1.emf, as(emf) replace
graph combine bit3.gph bit4.gph, title(Ranking of 1960s HEIs by UoA rankings)
graph export $path2\hei_60s_rank1.emf, as(emf) replace
* Example of unit level data from Stirling Univ.
list uoa3 s4 s3 fte gpa* unit_g rusmean if hei3==139
*****************
```

Panel 2: Illustrative unit level data from Univeristy of Stirling

	uoa3	s4	s3	fte	gpa_1	gpa_2	gpa_2u	unit_g	rusmean
2075.	Computer S	5	40	10	2.4	37.50419	46.13223	48.43189	.4217909
2076.	European S	5	25	12.03	2.05	44.66712	46.13223	46.62803	.2327407
2077.	Economics	15	45	7.6	2.75	37.97216	46.13223	54.07967	.5294637
2078.	Agricultur	5	45	32.9	2.45	52.27324	46.13223	48.49041	.5301701
2079.	Politics a	5	10	6	1.75	28.79008	46.13223	51.05609	.4819652
2080.	Philosophy	25	45	11	2.95	59.95757	46.13223	54.21309	.6516607
2081.	Law	5	35	7	2.15	38.41819	46.13223	50.5577	.507668
2082.	History	15	35	18.5	2.5	41.84967	46.13223	49.75456	.4476549
2083.	Sports-Rel	15	25	11.6	2.4	60.20428	46.13223	41.31431	.0973163
2084.	Accounting	0	45	10.8	2.35	50.27791	46.13223	47.58783	.1629073
2085.	Education	15	40	20.6	2.6	59.36917	46.13223	48.74143	.351389
2086.	Communicat	10	60	14	2.75	55.04335	46.13223	40.71224	.1259518
2087.	Nursing an	20	30	16.6	2.5	50.12636	46.13223	43.5994	.27137
2088.	Social Wor	10	45	25.8	2.6	50.89356	46.13223	46.76595	.3135845
2089.	Earth Syst	5	35	19.8	2.35	30.83045	46.13223	49.32261	.4820471
2090.	Psychology	5	10	24.2	1.9	34.69257	46.13223	47.41619	.355085
2091.	Business a	10	30	31.9	2.3	41.16507	46.13223	48.90034	.3532122
2092.	English La	10	45	17	2.55	45.70628	46.13223	47.39708	.3829423

Key: As table 2, panel 4, plus: gpa_2 = unit level gpa standardised within UoA average; gpa_2u = HEI level average of gpa_2; unit_g = UoA level average of host HEI's gpa_2; rusmean = UoA level average of proportion of host HEI's within Russell group.

When the measures derived in Table 8 are presented, we see examples of variation in the apparent ranking of different UoA's. The example of Stirling University is again instructive, because we now see that within the relatively high achieving submissions, some units were within UoA's which themselves had relatively high research profiles according to the 'unit_g' and 'rusmean' indicators (such as Philosophy), but others were found in UoA's with apparently lower relative research standards (such as Sports Studies, Communications and media, Nursing, and Social Work). These patterns are increasingly different from the overview data obtained from inspection of unstandardised proportions, and, we argue, should be highly relevant to policy decisions made upon the basis of the RAE.

2.8 Category scorings

Our next analysis turns to the final problem with summarising categorical data introduced above. We have already described the widely used 'Grade Point Average' measure which summarises the arithmetic average of numeric values assigned to the five RAE categories. A single parameter summary of this measure (i.e. its average) within or across HIE's and UoAs is clearly attractive to communicating the results of analysis. However numerous methodologists have warned against simplistic arithmetic summaries of categorical divisions. At the very least, it seems hard to understand why the numerical labels 4, 3, 2, 1 and 0 should necessarily be the optimum arithmetic levels for summarising the RAE grading decisions.

The relative impact of variations in these numerical equivalences does in fact turn out to be generally slight (this is unfortunate, because the possible alternatives are viewed with considerable lay suspicion, and typically assumed to constitute devices by which statistical evidence can be manipulated). Nevertheless, it is highly relevant to try out several alternatives, a procedure that it rarely followed in existing research projects.

An attraction of the Stata format workflow model built up above is that it is a relatively simple procedure to replicate analyses with a different formulation for an outcome variable. Thus, all the comparisons from sections 2.2 to 2.7 above, undertaken on Grade Point Average results, can be rapidly repeated for an alternative derivation of scores.

The question then arises of which alternative numerical measures (functional forms) to consider. The existing scale points might perhaps be squared or cubed to increase the relative influence of the higher gradings. We might take a priori decisions to use only one of the criteria as our crucial marker (e.g. the proportions achieving 4* grades contrasted with all others). Lastly, a more attractive empirical approach would be to actively investigate the arithmetic structure of responses, and assign numerical equivalents accordingly. For example, a stereotyped ordered logistic regression model could be employed to assign category scores on the basis of patterns of correlations with other criteria.

To keep our demonstration simple, in table 9 we show one replication of the analyses above (from section 2.6 – cf. Table 6) with a new arithmetic averaging function for GPA – namely, assigning scores of 100 for 4*; 50 for 3*; 20 for 2*; 0 for 1*; and -40 for uc. Readers may note that the only difference between the two Stata extracts [Stata-6] and [Stata-9] is in the line beginning 'gen gpa_1' – this leads to a relatively

rapid and easily conducted replication. Substantively, readers may also note that, for the example of Stirling, reorienting the summary measure in this way (which, arithmetically, places greater influence on the extreme positive and negative values) has negligible effect on the relative rankings within the HIE and in each unit's relative position within its own UOA.

Table 9: Grade Point Average standardisation within UoA's

Panel 1: [Stata-9]

```
******** [Stata-6]
** Mean standardisations using conventional gpa-measure
use $path1\rae2008_2.dta, clear
capture drop gpa_1
gen gpa_1 = (s4*100 + s3*50 + s2*20 + uc*-40) / 100
sav $path9\m1.dta, replace
summarize gpa_1
collapse (mean) mean1=gpa_1 (sd) sd1=gpa_1 [aweight=fte], by(uoa3)
summarize
sort uoa3
sav $path9\m2.dta, replace
use $path9\m1.dta, clear
sort uoa3
merge uoa3 using $path9\m2.dta
tab _merge
drop _merge
capture drop gpa_2
gen gpa_2 = 50 + (15*((gpa_1 - mean1) / sd1))
summarize gpa_2 [aweight=fte]
* Example: standardised subject rankings at Stirling university
list uoa3 s4 s3 s2 s1 fte gpa* if hei3==139
table uoa3 if hei3==139, c(mean s4 mean s3 mean fte mean gpa_1 mean gpa_2) format(%4.2f)
* Example: ranking within 1960 univs using within uoa zscore graph hbar (mean) gpa_2 if
sixties==1 [aweight=fte], ///
 over(hei3, label(labsize(vsmall)) sort(1) ) ///
    title(1960's: HIE mean gpa's standardised within UoAs)
```

Panel 2: Revised gpa and within-UoA standardised GPA for University of Stirling

uoa3	mean(s4)	mean(s3)	mean(fte)	__000005	__000006
Accounting and Finance	0.00	45.00	10.80	31.50	47.20
Agriculture, Veterinary and Food Science	5.00	45.00	32.90	35.50	50.48
Business and Management Studies	10.00	30.00	31.90	33.00	41.28
Communication, Cultural and Media Studie	10.00	60.00	14.00	45.00	52.21
Computer Science and Informatics	5.00	40.00	10.00	34.00	35.59
Earth Systems and Environmental Sciences	5.00	35.00	19.80	32.50	30.37
Economics and Econometrics	15.00	45.00	7.60	45.50	37.53
Education	15.00	40.00	20.60	42.00	57.79
English Language and Literature	10.00	45.00	17.00	39.50	43.21
European Studies	5.00	25.00	12.03	24.50	43.43
History	15.00	35.00	18.50	38.50	40.42
Law	5.00	35.00	7.00	28.50	38.60
Nursing and Midwifery	20.00	30.00	16.60	40.00	49.85
Philosophy	25.00	45.00	11.00	53.50	58.24
Politics and International Studies	5.00	10.00	6.00	17.00	30.10
Psychology	5.00	10.00	24.20	21.00	35.12
Social Work and Social Policy & Administ	10.00	45.00	25.80	40.50	48.70
Sports-Related Studies	15.00	25.00	11.60	36.50	59.42

Key: As table 2, panel 4, plus: _00005 = revised numeric average of RAE categories; _000006 = unit level version of revised average standardised according to UoA averages (to mean 50, sd 15).

2.9 Comparability with the 2001 RAE exercise

The UK's Research Assessment Exercises are long, slow processes. The last RAE was published in 2001, covering research outputs between 1996 and 2001. The ranking criteria used in that RAE became widely understood throughout the Higher Education sector up to the present period, and so a natural question arises as to how the 2001 and 2008 results can be meaningfully compared.

In the case of the RAE, it might be argued that the differing methodologies and category criteria between the exercises mean that the 2001 and 2008 data simply cannot be compared. Alternatively, it might also be advocated that the category criteria of the two exercises are equivalent and therefore the proportions of members of UoAs achieving relevant criteria may be directly compared. Both of these positions would be consistent with an approach to harmonisation which insists upon 'identity equivalence' before making comparisons (cf. van Deth, 2003). In this perspective, comparisons may or may not be made depending upon a reader's willingness to believe in the consistency of rating criteria over time, and their willingness to accept the irrelevance of other changes in the underlying distribution of grades. Because the identity equivalence perspective requires no significant further analytical work, it is commonly adopted in making comparisons between RAE data.

However, in our view the RAE is primarily about making comparisons over UoA's relative positions within the distribution of ratings at each period. Such an interest is not well suited to an approach of 'identity equivalence' for comparative evaluations, but is more suitable to the concept of 'meaning equivalence' (used by van Deth, 2003 to describe the use of measures and statistical summaries which carry the same relative meaning between contexts). Such a view would hold that changes in the underlying distribution of ratings (such as the process of 'grade inflation' in educational qualifications) should impact upon the account of RAE results; and that the same (or nominally the same) rating criteria over different RAE's might not necessarily correspond to the same performance over time.

Achieving comparability across time (or between other contexts such as in cross-national research) is a major objective of many standardisation procedures. It is ultimately achieved through activities of 'data management' (and, we would argue, is under-used in many comparative analyses precisely because of the perceived extra analytical work involved).

In Table 10 we present SPSS and Stata syntax illustrating one sequence of commands in data management which allowed us to create a new database which illustrates thresholds for the 2008 GPA score, *within UoAs*, which we argue correspond to the category thresholds used in the 2001 RAE. The net result is a database organised by UoA which lists different thresholds for different UoA's. This calculation is achieved by (1) calculating the proportions of submitted staff in each UoA achieving the 2001 thresholds, then (2) identifying the proportions of submitted staff achieving the same relative positions in the 2008 RAE. This exercise means, by definition, that a UoA cannot change its overall profile between the two RAE's, which may be unsatisfactory assumption (particularly for UoA's whose composition has changed

significantly over time). It does however control for changes within the size of different HEI units within each UoA.

Table 10: Linking 2008 RAE results with 2001 result categories

Panel 1: [SPSS-2]

```
********** [SPSS-2] : Access 2001 data; conversion to SPSS and Stata data  :.
* (2001 data from http://www.hero.ac.uk/rae/Results/ ).
get data /type=xls   /file=!path1+"all1.xls"
   /sheet=name "simple2001"
   /cellrange=range 'a1:i2599'  /readnames=on  .
descriptives var=all.
*.
autorecode var=hei heiname uoa01name rating prop /into=hei2 hei3 uoa01_3 rating2 prop2.
descriptives var=hei2 hei3 uoa01 uoa01_3 rating2 prop2 fte .
fre var=rating2 prop2 .
descriptives var=all.
sav out=!path1+"rae2001_1.sav".
sav translate out=!path1+"rae2001_1.dta"
   /type=stata /version=8 /replace.
**********.
```

Panel 2: [Stata-10]

```
*** [Stata-11]: Linkage between RAE2001 and RAE2008 ratings

* 1) Work using 2001 data derived from [SPSS-2]
use $path1\rae2001_1.dta, clear
numlabel _all, add
tab uoa01_3
* i) Derive an indicator of UoA's felt to be equivalent across time
gen huoa=uoa01
global uoa01var "huoa"
global uoa08var "Nothing"
do "http://www.dames.org.uk/rae2008/uoa0108recode.do"
label variable huoa "Unit of Assessment in 2001-2008 harmonised coding"
summarize huoa uoa01
tab huoa
* ii) Calculate thresholds for category rankings in the harmonised coding
tab rating2
tab rating2, gen(level)
collapse (mean) level* [aweight=fte], by(huoa)
summarize
replace level2=level1+level2
replace level3=level3+level2
replace level4=level4+level3
replace level5=level5+level4
replace level6=level6+level5
replace level7=level7+level6
summarize level*
sort huoa
save $path9\t1.dta, replace

* 2) Work using 2008 data
* iii) Open 2008 data and derive an indicator of UoA's felt to be eqivalent across time
use $path1\rae2008_2.dta, clear
gen huoa=uoa
global uoa01var "Nothing"
global uoa08var "huoa"
tab $uoa08var
do "http://www.dames.org.uk/rae2008/uoa0108recode.do"
label variable huoa "Unit of Assessment in 2001-2008 harmonised coding"
summarize huoa uoa
tab huoa

* iv) link with thresholds
sort huoa
merge huoa using $path9\t1.dta
tab _merge
keep if _merge==1 | _merge==3
drop _merge
* (comment - we now have 2008 unit level data linked with 2001 within uoa category thresholds)
```

```
* v) Derive weighted relative rankings (according to standard gpa measure) within
*    harmonised units of assessment :
capture drop gpa_1
gen gpa_1 = (s4*4 + s3*3 + s2*2 + s1*1) / 100
gen power_1 = gpa_1*fte
sav $path9\m1.dta, replace
collapse (mean) mean1=gpa_1 (sd) sd1=gpa_1 [aweight=fte], by(uoa3)
sort uoa3
sav $path9\m2.dta, replace
use $path9\m1.dta, clear
sort uoa3
merge uoa3 using $path9\m2.dta
drop _merge
capture drop gpa_2
gen gpa_2 = 50 + (15*((gpa_1 - mean1) / sd1))
summarize gpa_2 [aweight=fte]
* (unweighted rankings and percentiles)
capture drop rank1
egen rank1=rank(gpa_1) , by(huoa)
capture drop ucount
egen ucount=count(gpa_1), by(huoa)
capture drop pct1
gen pct1= rank1 / (ucount+1)
summarize rank1 ucount pct1
gen uw2001=-9
replace uw2001=1 if pct1 >= 0
replace uw2001=2 if pct1 >= level1
replace uw2001=3 if pct1 >= level2
replace uw2001=4 if pct1 >= level3
replace uw2001=5 if pct1 >= level4
replace uw2001=6 if pct1 >= level5
replace uw2001=7 if pct1 >= level6
label define l2001 1 "Grade 1" 2 "Grade 2" 3 "Grade 3a" 4 "Grade 3b" 5 "Grade 4" 6  "Grade 5" 7 "5 star"
label values uw2001 l2001
numlabel _all, add
tab uw2001
list uoa3 gpa_1 uw2001 if hei3==139
* (weighted rankings and percentiles)
capture drop sumfte
egen sumfte=sum(fte), by(huoa)
gsort +huoa +rank
bysort huoa (rank1): gen csum=sum(fte)
* (above using tip from Nick Cox at: http://www.stata.com/statalist/archive/2002-07/msg00160.html)
list huoa rank fte sumfte csum in 1/80
capture drop pct2
gen pct2= csum / (sumfte+1)
summarize rank1 ucount csum sumfte pct2
gen w2001=-9
replace w2001=1 if pct2 >= 0
replace w2001=2 if pct2 >= level1
replace w2001=3 if pct2 >= level2
replace w2001=4 if pct2 >= level3
replace w2001=5 if pct2 >= level4
replace w2001=6 if pct2 >= level5
replace w2001=7 if pct2 >= level6
label values w2001 l2001
numlabel _all, add
tab w2001
numlabel _all, remove
* Example data for University of Stirling, using 2001 categories and distributions
gsort -pct2
list uoa3 gpa_1 gpa_2 pct2 w2001 if hei3==139
* Example data for Nursing and Midwifery panel, using 2001 categories and distributions
list hei3 gpa_1 gpa_2 pct2 w2001 if uoa3==49

** 3) Reformating and exporting for publication at www.dames.org.uk:
keep hei3 uoa3 fte uc s1 s2 s3 s4 gpa_1 gpa_2 power_1 rank1 pct2 w2001 ///
      hei heiname uoa uoaname huoa mult1 mult2 fte
summarize
describe
label variable hei3 "Higher education institution, numeric coding"
label variable uoa3 "Unit of assessment, numeric coding"
label variable fte "Number of full time equivalent staff within unit"
label variable huoa "Unit of assessment, 2001-2008 linkage variable"
label variable s1 "Proportion graded 1 star"
```

```
label variable s2 "Proportion graded 2 star"
label variable s3 "Proportion graded 3 star"
label variable s4 "Proportion graded 4 star"
label variable uc "Proportion graded unclassified"
label variable gpa_1 "Conventional arithmetic Grade Point Average"
label variable gpa_2 "Conventional GPA standardised within UoA"
label variable power_1 "Conventional GPA times FTE"
label variable rank1 "Relative ranking within UoA"
label variable pct2 "Weighted percentile ranking within UoA"
label variable w2001 "RAE result in 2001 units, using weighted percentile 2008 ranks on 2001 thresholds"
summarize
numlabel _all, add
saveold $path1\rae2008_3.dta, replace
dir $path1\*.dta
```

Panel 3: Illustrative results for the University of Stirling

	uoa3	gpa_1	gpa_2	pct2	w2001
251.	Philosophy	2.95	59.95757	.8457449	5 star
348.	Sports-Related Studies	2.4	60.20428	.7882057	5 star
491.	Education	2.6	59.36917	.6991579	Grade 4
593.	Communication, Cultural and Media Studies	2.75	55.04335	.6364117	Grade 4
625.	Nursing and Midwifery	2.5	50.12636	.6147168	Grade 3b
690.	Agriculture, Veterinary and Food Science	2.45	52.27324	.582757	Grade 5
797.	Accounting and Finance	2.35	50.27791	.5298879	Grade 5
904.	Social Work and Social Policy & Administration	2.6	50.89356	.4722171	Grade 4
1111.	European Studies	2.05	44.66712	.3723886	Grade 4
1222.	English Language and Literature	2.55	45.70628	.3268699	Grade 4
1390.	Business and Management Studies	2.3	41.16507	.2587766	Grade 3a
1431.	History	2.5	41.84967	.2440363	Grade 4
1476.	Computer Science and Informatics	2.4	37.50419	.228477	Grade 4
1631.	Psychology	1.9	34.69257	.1745268	Grade 3b
1678.	Economics and Econometrics	2.75	37.97216	.1619764	Grade 3b
1720.	Law	2.15	38.41819	.146867	Grade 4
1725.	Earth Systems and Environmental Sciences	2.35	30.83045	.1450686	Grade 3a
2014.	Politics and International Studies	1.75	28.79008	.0688711	Grade 3a

Key: As Table 2, panel 4, plus: gpa_2 = within UOA standardised GPA to mean 50, sd 15; pct2 = percentile ranking of staff within UoA; w2001 = 2001 RAE categories according to percentile thresholds for FTE staff submitted within the 2001 RAE.

Panel 1: [SPSS-3]

```
****************** [SPSS-3] : Opens Stata format data from [Stata-10],
*    exports to SPSS and MS Excel format for www.dames.org.uk/rae2008/
get stata file=!path1+"rae2008_3.dta"   .
descriptives var=all.
sort cases by uoa3 gpa_2 power_1 .
sav out=!path1+"rae2008_3.sav".
sav translate out=!path1+"rae2008_3.xls"
  /type=xls /fieldnames /replace.
**********.
```

The results generated and presented in Table 10 are important since they are (to our knowledge) the only published databases listing 2008 RAE results in terms of 2001 criteria employing an approach of 'meaning equivalence'. For the readers' convenience, we have also placed the derived data files in SPSS, Stata and MS Excel format available on our website (www.dames.org.uk/rae2008/ for data files rae2008_3.[dta/xls/sav] respectively).

The third panel of Table 10 lists illustrative results of this exercise for the units from the units submitted from the University of Stirling. We see that when mapped to the 2001 categories and using the 2001 percentile thresholds, some UoAs have more or less advantaged relative positions than when considered using unstandardised 2008 rankings. To repeat, this particular method forces each UoA panel to have the same proportions of FTE staff within the same categories as the 2001 exercises, and so it may misrepresent individual units within panels where composition has significantly

changed, or where overall research standards have improved or fallen significantly between 2001 and 2008. Nevertheless, we consider that this mapping of 2008 results to 2001 categories is a particularly effective way to understand the 2008 RAE rankings.

2.10 Summary

In summary, by presenting these detailed ('low level') accounts of data management operations and simple approaches to standardising categorical data, we have used the examples from Sections 2.2 to 2.9 to illustrate numerous extension analyses of the existing RAE data which can readily be performed but are not, to the best of our knowledge, widely considered, despite the significant structural consequences potentially associated with the RAE results.

3. Summary Analysis of the UK 2008 RAE results

For practical reasons we present below a first draft summary of what we consider to be key comparative results from the UK RAE that have not previously been disseminated. We first present graphical summaries of the results, and summarise their findings with five key arguments about RAE result which (to our knowledge) conflict with emphases previously published on this topic. We intend to expand upon these analyses and present further discussion in later editions of this draft paper.

Figure 1.

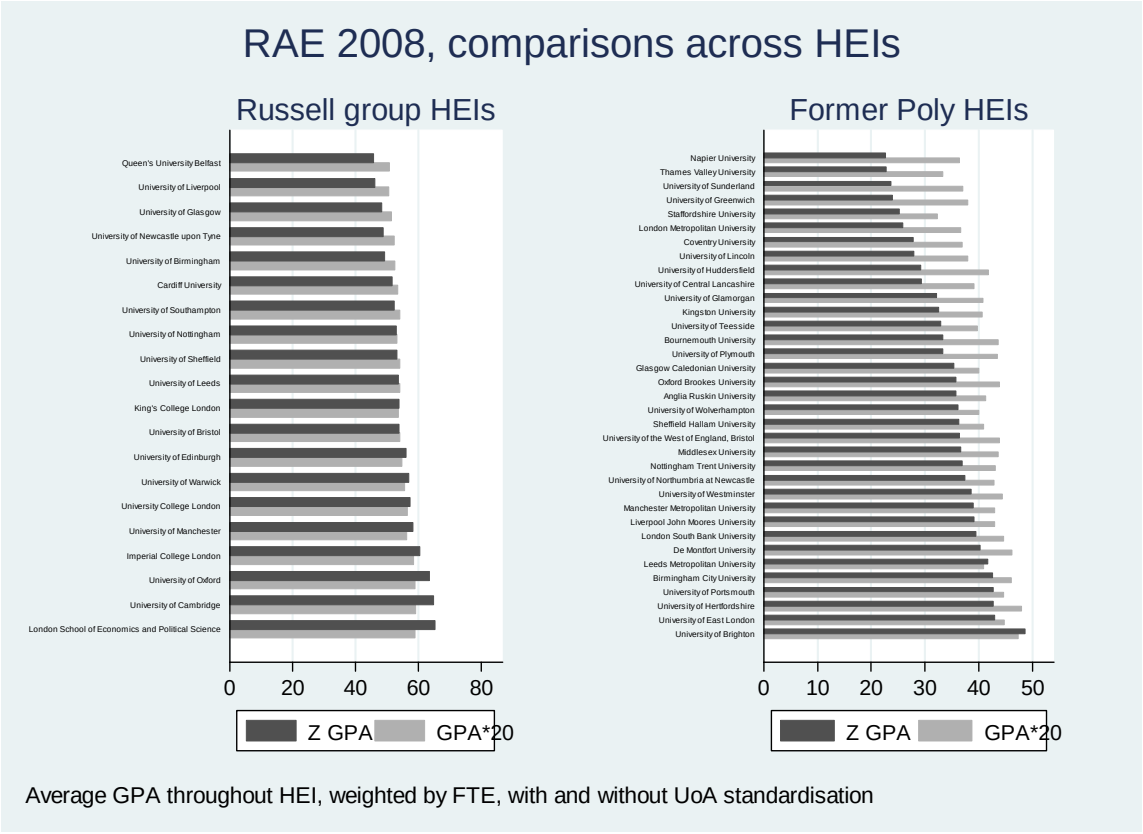
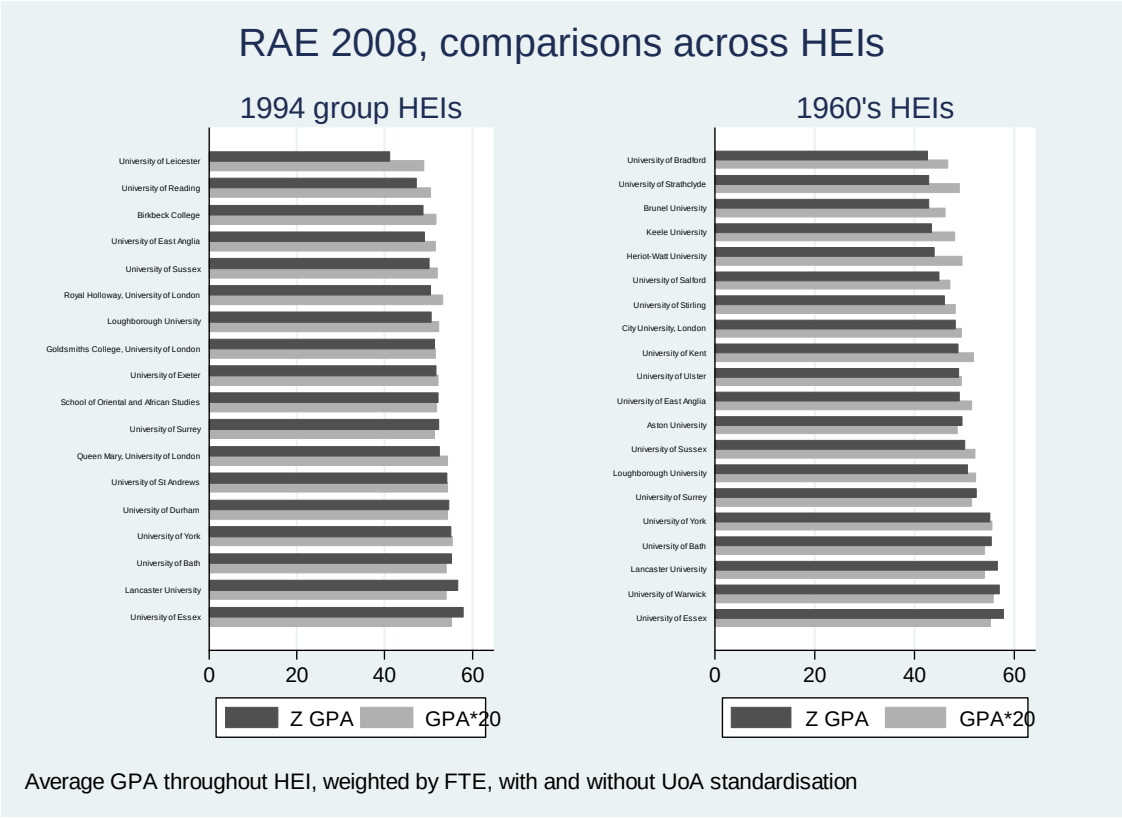


Figure 2.



Average GPA throughout HEI, weighted by FTE, with and without UoA standardisation

Figure 3.

RAE 2008: average zGPAs of HEIs hosting UOA staff (1), and percent of UOA staff in Russell group institutions (2)

- **LSE outranks Oxbridge.** As explained above our general preference for measures summarizing an HEI's performance across different UoA's is to use of a measure which takes some steps towards standardisation. The average summaries of relative position within the respective UoAs (the darker bars from Figure 1, by which the HEI's are ranked) places LSE above the Oxbridge institutions, a position contrary to the unstandardised rankings widely reported in press coverage

- **The RAE reinforces major sectoral cleavages between groups of HEIs**. Summarising RAE statistics across different groups of HEIs makes clear the substantial levels of difference between HEI groups (see also Section 2.6 above). The Russell group institutions have the strongest average research profiles, and groups such as the former polytechnics much lower average rankings. It is often informative to note HEI rankings within institutional groups – for instance, as shown in Figure 2, University of Essex is a consistently high achiever within the cohort of Universities which received their royal charters in the 1960's.

- **Higher performances within non-Russell group institutions may often be over-exaggerated.** Many HEI's outside the handful with the highest average RAE results still presented optimistic accounts of their RAE performances. However, as Figure 1 highlights, performance in many of the non-traditional sectors tends to benefit disproportionately from performance on non-standardised measures. Moreover, as Figure 3 highlights, UoAs which may have relatively lower research standards tend to be more commonly found outside the Russell group.

- **RAE grade point averages may under-emphasise disparities in research quality.** The standard GPA measure, and other minor variations achieved by changes in numerical equivalence) tend to have low degree of spread amongst major HEIs, and an implausibly low degree of spread between UoAs. For standard GPA measures, the source of spread frequently lies with extreme low or high values of a specialist institution. Using standard measures, measures of 'research power' offer much more discrimination between institutions and departments, and may be more appropriate. In addition, comparisons of UoA profiles (Figure 3) suggest that UoAs themselves differ in their overall profiles to a greater extent than is revealed by unstandardised measures, and in a way which is structured according to HEI types.

- **Many apparent improvements in individual performances in the 2008 RAE would not persist if evaluated according to RAE 2001 distributions.** As shown in the results presented in section 2.10, adjusting 2008 results to the 2001 thresholds often has the effect of degrading upward movements. The explanation may be that it is commonly the case that higher achieving units have also expanded the size of their FTE submission between 2001 and 2008. Therefore, apparent upward movements on GPA scores achieved by smaller Units remain insufficient to ensure a comparable upward movement within the percentile distribution of all FTE staff within the UoA.

4. Conclusions

In this short paper we have sought to make two methodological points. Firstly, that a small effort in 'data management' (which is readily accommodated within mainstream database software such as SPSS and Stata), can make a substantial difference to the results of analysis of quantitative data. We have tried to show that data manipulations such as the calculation of weighted, standardised scores, and deterministic data linkages with external resources (as in our example of linking with a record of typologies of HEIs), can give revealing insights about how the RAE 2008 results vary across locations. Secondly, we have tried to demonstrate how numerous alternative strategies to analysing categorical data can readily be applied to records such as the RAE databases, and that these approaches can have some impact on results (though in some instances, concerning the scoring of RAE categories, perhaps not as much impact as is popularly thought).

This paper is motivated by the methodological attention to data management tasks of the DAMES research Node, an initiative in 'e-Social Science'. The contribution of e-Science in this field is not with the capacity to perform the relatively simple manipulations and analyses presented (as shown above, these can be readily achieved in packages such as SPSS and Stata). Rather, the contribution of e-Science is as a wider approach to organising data and analysis – such as in setting standards for distributing data and resources (and metadata describing them) in order to make it easier for researchers to 'discover' and exploit suitable data to enhance their analysis; and in modelling effective workflows of intricate data management operations in order to maximise the consistency and replicability of research analyses. Over the period 2008-2011 the DAMES Node is developing and delivering services for social science data, which will offer easier facility in the type of data management and analysis activities described above, to the mainstream social science research community.

A rapidly emerging theme of the DAMES Node is the need to encourage analysts to 'do something, not nothing' in the field of data management and manipulations. By this, we mean that we see many scenarios in social science research where a relatively light input in the areas of data linkage and variable manipulations could pay dividends for analytical aims. Nevertheless, we also witness a great deal of reluctance amongst many researchers to take such steps – arising, we speculate, from an irrational fear of the complexities and errors likely to occur when running such routines. We hope with this short example using UK RAE data that we have demonstrated how tractable many relevant, apparently complex, data management techniques can in fact be.

References

Blossfeld, H. P., Grolsch, K., & Rohwer, G. (2007). *Event History Analysis with Stata*. New York: Lawrence Erlbaum

Bottero, W. (2005). Interaction distance and the social meaning of occupations. *Sociological Review, 53*(2), 56-72.

Caulkin, S. (2008, 21 December). A senseless system graduates without honours *The Observer*.

guardian.co.uk. (2008). RAE 2008: Results for UK universities. Retrieved 18 December 2008, from http://www.guardian.co.uk/education/table/2008/dec/18/rae-2008-results-uk-universities

HESA. (2008). HESA - The Higher Education Statistics Agency. Retrieved 1 December 2008, from http://www.hesa.ac.uk/

Kohler, H. P., & Kreuter, F. (2009). *Data Analysis Using Stata, 2nd Edition*. College Station, Texas: Stata Press.

Levesque, R., & SPSS Inc. (2008). *Programming and Data Management for SPSS Statistics 17.0*. Chicago, Il.: SPSS Inc. .

Lipsset, A. (2008a, 18 December 2008). Nursing research takes its place on world stage: Departments around the country score strongly in a turnaround that reflects the higher academic status of nurses. *The Guardian*.

Lipsset, A. (2008b, 18 December). RAE: The 'dismal science' is cause for celebration. *The Guardian*.

Morgan, S. L., & Winship, C. (2007). *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge: Cambridge University Press.

RAE. (2008). RAE 2008: Research Assessment Exercise. Retrieved 18 December, from http://www.rae.ac.uk/

Scott Long, J. (2008). *The Workflow of Data Analysis Using Stata*. Boca Raton: CRC Press.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103*, 677-680.

van Deth, J. W. (2003). Using Published Survey Data. In J. A. Harkness, F. J. R. van de Vijver & P. P. Mohler (Eds.), *Cross-Cultural Survey Methods* (pp. 329-346). New York: Wiley.