



Communicating knowledge: How and why researchers publish and disseminate their findings

Supporting paper 1: Bibliometric analysis

Jenny Fry; Charles Oppenheim - DIS, Loughborough University

Claire Creaser; William Johnson; Mark Summers; Sonya White - LISU, Loughborough University

Geoff Butters; Jenny Craven; Jill Griffiths; Dick Hartley - CERLIM, Manchester Metropolitan University

September 2009



Table of Contents

Executive summary	iii
Key findings.....	iii
1 Introduction	1
1.1 Scope of study	2
2 Methodology.....	3
2.1 Sampling frame	3
2.2 Sampling process.....	3
2.3 Obtaining lists of outputs.....	4
2.4 Collection of materials and categorisation of bibliographical references	5
2.4.1 WoS sourced material	5
2.4.2 Non-WoS sourced material	6
2.4.3 Categorisation of references	6
2.5 Challenges	6
2.5.1 Identifying the correct researcher	6
2.5.2 Compiling accurate reference lists	6
2.5.3 Researchers with no identifiable dissemination in the appropriate year.....	7
2.5.4 Dissemination without citation	7
2.5.5 Inability to locate an accessible copy of outputs.....	7
2.5.6 Reference counting	8
2.5.7 Referencing systems	8
2.6 Notes on the analysis.....	8
3 Results	10
3.1 Types of dissemination	10
3.2 Collaboration	11
3.3 Citation practice	15
4 Discussion and conclusions	20
4.1 Recommendations concerning methodology	21
Appendix: Allocation of subjects to broad disciplines	23
Medical sciences	23
Physical sciences	23
Engineering	23
Social sciences.....	24
Humanities	24
Arts	24
Education	24

Acknowledgements

The research on which this report is based was undertaken by a team from Loughborough University and Manchester Metropolitan University.

The research team would like to thank:

- Aaron Griffiths and Michael Jubb from the Research Information Network
- The Expert Panel - Michael Anderson; Bob Campbell; Hannah Chaplin; Alison Holt; Neil Jacobs
- All those researchers who gave their time to attend and contribute to the focus groups, and who completed the survey
- Mary Ashworth & Sharon Fletcher from LISU, who provided invaluable administrative support
- The bibliometric data collectors – Karen Davies; Tracy Forskitt; Vicki Jackson; Amy Beeston
- The focus group data collectors – Evgenia Vassilakaki, Magda Vassiliou and Ioanna Zorba

Executive summary

Methods

- Two years' data were selected for analysis, from separate research assessment periods – 2003 and 2008
- Two samples of authors were drawn from RAE submissions for 2001 and 2008, and lists obtained of their published research outputs in 2003 and 2008 respectively
- Between 40% and 50% of authors sampled for each year had no identifiable research outputs in the relevant year
- Outputs identified were examined, and their references categorised and counted
- A total of 1,452 works from 484 authors were included in the analysis

Key findings

Dissemination practice

- There were significantly more outputs per author in 2008 than in 2003, particularly in Biomedicine, and social sciences
- There were significantly more journal articles, and fewer monographs, in 2008 than in 2003
- There were significantly more multi-authored works in 2008 than in 2003, particularly in social sciences and physical sciences
- There were significantly more inter-institutional collaborations, and more international collaborations, in 2008 than in 2003

Citation practice

- There was no difference in the average number of citations per output between 2003 and 2008 overall
- There were significant differences between disciplines in the numbers of citations per publication – humanities cite the greatest number of works on average; engineering the fewest
- Monographs had an average of over 230 references each, compared to 38 for journal articles and 47 for book chapters
- Significantly more journal articles, and fewer books and grey literature works were cited in 2008 than in 2003
- Biomedicine, physical sciences and social sciences cite twice as many articles per publication as other disciplines
- Humanities, and, to a lesser extent, social sciences and education, cite more books per output on average than other disciplines
- Social sciences, and, to a lesser extent, education, cite more grey literature per output on average than other disciplines
- Social sciences, education and humanities cite more websites than Biomedicine, physical sciences or engineering
- Books and book chapters are most likely to cite books/book chapters
- Conference outputs are most likely to be cited in conference proceedings

1 Introduction

This study was commissioned by the Research Information Network (RIN), in collaboration with the Joint Information Systems Committee (JISC), in December 2008, to gather and analyse evidence about:

- The motivations, incentives and constraints that lead researchers in the UK in different subjects and disciplines to publish and disseminate their work in different ways
- How and why researchers cite other researchers' work
- In particular, how researchers' decisions on publication and citation are influenced (or not) by considerations arising from research assessment

The following key issues were investigated, covering three broad areas:

1. Publication and dissemination behaviour:

- What factors make a scholar prefer one dissemination medium over another?
- Has the trend towards electronic publishing, and open access to scholarly outputs of all types affected their preferences, and how?
- To what extent does collaboration across institutions or disciplines affect these behaviours?
- What motivates scholars to publish, and what constrains them?
- What impact will the advent of electronic publishing have in the future on scholars' dissemination activities?

2. Citation behaviour:

- Why do scholars cite the way they do?
- Are scholars' reading behaviours, and therefore their knowledge of the prior literature, changing?

3. The perceived influence of research assessment (past and anticipated):

- What pressures are scholars feeling regarding dissemination because of the increasing importance of the RAE/REF, and how is that pressure manifesting?
- Have pressures from the RAE/REF affected publishing behaviours?

Four complementary methodologies were used, and the detailed methods and outcomes are described in a series of four supporting papers to the main project report. These supporting papers are all available at www.rin.ac.uk/communicating-knowledge:

1. Bibliometric analysis (this document)
2. Report of focus group findings
3. Report and analysis of researcher survey
4. Literature review

This mixed methods approach has been used to help ensure an holistic view is provided of the publication, dissemination and citation behaviour of researchers across subject disciplines. It will also establish a baseline for further studies in this area. Consequently, the individual reports should not be read in isolation, but in conjunction with the main RIN project report [*Communicating knowledge: How and why UK researchers publish and disseminate their findings*](#) (September 2009). Each individual report provides further detail and supporting evidence for the material presented in the main report.

1.1 Scope of study

The research team wanted to compare the behaviour of researchers when publishing and disseminating their research as reported in the focus groups and survey, with behaviours which are evidenced by published research outputs. In order to do this, a bibliometric analysis was carried out, focussing on examination of a sample of research outputs and the material cited therein. Data were collected on the numbers and types of output produced, and numbers and types of outputs cited in the published research. A random sampling design was used, so that inferences could be made about the national picture, and broad comparisons drawn between disciplines, and over time.

The analysis faced a number of challenges, not least the resource intensive nature of collecting data at this level of detail. The ideal for this type of analysis would be to look at a five or even ten year trend, analysing data from each year, but the timescale of this project, and available resources, did not allow this. As a result, a pragmatic approach was adopted, restricting the analysis to two snapshot output years. A number of options were considered to select the two years, bearing in mind the objective of the overall project to investigate the influence of research assessment. While it might seem preferable to analyse data from years in the middle of each research assessment period – i.e. 2004 from the 2008 Research assessment Exercise (RAE) – the relevant year from the 2001 RAE (1997) was thought potentially too distant to find information on non-electronic and non-peer reviewed outputs with any degree of consistency. The team therefore started by selecting 2008, the most recent complete year available, which will be included in the research assessment period after the 2008 RAE (which took into account outputs before 31-12-07). 2008 also has the advantage of corresponding more closely to the survey and focus groups, which collected current data and opinions. To give a comparison over time, 2003 was selected. It was thought that this would be a sufficient period to establish differences, but not so long ago that it would be difficult to find outputs.

The intention was that the method should be replicable for the outputs from any period, and as such it has been fully documented below.

2 Methods

In outline, the method adopted for this study was to:

- Devise a sampling frame of UK academics with outputs in 2003 and 2008;
- Draw a stratified random sample from each list;
- Obtain details of the published outputs for that year for each member of the sample;
- Obtain, examine and categorise the items in the bibliography of each output.

2.1 Sampling frame

It was initially proposed to compile a sampling frame of UK academics from data in the Web of Science (WoS). However, in attempting to do this it was found that, despite using an automated script to parse the downloaded records in order to extract individual names from multiple-author articles, the resultant list contained an unhelpful mix of useful and irrelevant data.

An alternative approach was therefore adopted, utilising data from the RAE submissions in 2001 and 2008. These data sets provided a sampling frame which was pre-checked for UK higher education (HE) affiliation, and grouped by discipline (Unit of Assessment – UoA). The data source had the additional advantages that only the most research-active authors are submitted by each institution, so they can reasonably be expected to be producing research outputs on a regular basis. The dataset also covers a representative cross section of disciplines across UK HE. The disadvantage is that there is no guarantee that authors selected from this list would have outputs in 2003 or 2008. At the start of the project, data from the 2001 RAE were available in the public domain, but data for 2008 had not been released, and were sought from the Higher Education Funding Council for England (HEFCE).

2.2 Sampling process

The objective was to obtain a list of 400 names for each year of publication. Although using RAE returns to create the sampling frame removes the disciplinary bias inherent in WoS, a sampling design stratified by broad discipline was used, with the aim of producing a sufficiently large sample in the smaller discipline groups for comparative analysis. Weighting to reflect the proportions of academics in each broad discipline was applied in the analysis to give the picture for UK HE as a whole.

The sample stratification is illustrated in Figure 1. Appendix 1 lists the RAE Units of Assessment and HESA academic cost centres allocated to each broad discipline. Some of these allocations may appear arbitrary; the aim was to obtain lists of subjects within each broad discipline which would be comparable between the 2001 and 2008 RAEs, and consistent with the Higher Education Statistics Agency (HESA) academic cost centres, which provided data on the total numbers of academic staff. The stratification was informed by the distribution of academic staff returned in the 2001/2008 RAEs and the proportions of academic staff in post in 2006-07.

Figure 1 Sample design

	HESA ^a distribution	RAE 2001 distribution	RAE 2008 distribution	Average of these	Required sample size	Proportion of sample	Estimated weighting ^b
Bio-medicine	31%	24%	21%	25%	75	19%	1.35
Physical sciences	10%	15%	15%	13%	55	14%	0.96
Engineering	13%	12%	12%	12%	55	14%	0.89
Social sciences	21%	22%	26%	23%	75	19%	1.22
Humanities	10%	16%	16%	14%	60	15%	0.92
Arts	8%	6%	6%	7%	40	10%	0.67
Education	9%	5%	4%	6%	40	10%	0.58
	100%	100%	100%	100%	400	100%	

Notes:

^a Based on FTE academic staff 2006-07

^b The exact weighting to be used will depend on the actual sample achieved in each discipline area

Individuals who changed institution during the course of the RAE period may be entered twice in data set available, under each institution. In order to maintain the equivalence to simple random sampling within each discipline, suspected duplicate records (based on name and UoA) were removed, so that such individuals were not more likely to be selected than those who had not moved. The sampling frame was then sorted by UoA into broad discipline groups, and a systematic random sample (statistically equivalent to a simple random sample) drawn from each discipline group, using Excel's random number generator to set the starting point. A systematic sample was used to avoid any duplication in the sequence of random numbers generated by Excel.

The sample of authors to be used for the 2003 analysis was drawn at LISU based on the published 2001 RAE data. HEFCE provided the sample for 2008 following the LISU method. One issue with this approach is that the authors returned are those active in the period preceding each return, but there is no guarantee that authors selected for the sample will have outputs in the appropriate years. To compensate for this, an additional 10% was allowed in the sample. Initial investigations on the sample drawn for 2003 showed this allowance to be inadequate, and the sample was re-drawn, increasing the supplement to 30% of the desired sample size. The initial samples drawn were therefore: Biomedicine 98; sciences 72; engineering 72; Social sciences 98; humanities 78; arts 52; and education 52. In practice, even this larger sample proved insufficient to find 400 authors with outputs in the relevant year, but the project had insufficient time to further extend the samples.

2.3 Obtaining lists of outputs

Searches were undertaken to find all outputs from the sampled authors in each year. The searching was split into two sections: those subjects with good coverage in WoS (Biomedicine, science and engineering) and those with less or minimal coverage (social science, humanities, arts and education).

- WoS searches were performed using the selected author's name and target year and limited to those with a UK address.

It had been thought that the Directory of Open Access Journals (DOAJ) would provide additional bibliographic information, covering material not indexed by WoS. Searching for a sample group of researcher names yielded no additional useful results and, as a result, the DOAJ was not included in the searching procedure.

- For academics in subject areas not well covered by WoS, a variety of methods was used:
 - Researcher/departmental websites
 - Federated searches of a number of bibliographic databases, including IBSS, BHI, ArticleFirst, Communication Abstracts, Linguistics & Language Abstracts, Zetoc, Modern Languages Association International Bibliography, Art Full Text, OCLC Worldcat and ERIC.
 - Google Scholar and Google

Given the nature of the non-WoS subject disciplines, all forms of research output were sought. This is discussed further below.

Figure 2 summarises the achieved sample distribution across broad disciplines.

Figure 2 Summary of data collected

	2003				2008			
	Authors		Outputs		Authors		Outputs	
	Sample size	No with outputs	No. outputs analysed	No. references identified	Sample size	No with outputs	No. outputs analysed	No. references identified
Bio-medicine	97	45	102	102	96	52	235	235
Physical sciences	72	49	223	223	70	49	212	212
Engineering	72	27	87	87	70	44	149	149
Social sciences	98	37	75	109	96	52	137	166
Humanities	78	54	81	150	77	40	34	52
Arts	52	14	12	22	51	21	17	46
Education	52	26	51	63	51	22	37	63
	521	252	631	756	511	280	821	923

2.4 Collection of materials and categorisation of bibliographical references

This was carried out manually, primarily by information science students and recent graduates, supervised by members of the project team. A small sample of each individual's work was checked by members of the project team to ensure consistency of approach.

2.4.1 WoS sourced material

When relevant entries were identified in the search results, the indexed metadata and cited reference lists were printed and references categorised and counted.

2.4.2 Non-WoS sourced material

A list of items identified as being produced by the non-WoS group was drawn up for each year. A member of the project team then attempted to source a copy of the item, either in electronic form or in physical form from Loughborough University Library, another academic institution or the British Library. References were extracted from the material found, by printing or photocopying.

2.4.3 Categorisation of references

The cited items were categorised as follows:

- journal articles
- conference papers
- theses
- grey literature (e.g. technical reports; working papers; occasional papers; Governmental/NGO publications; British & international standards; mimeos)
- books (including book chapters)
- websites
- items in press or forthcoming
- other items (e.g. non-textual material; data sets; Parliamentary statutes; patents; historical documents; archive material)

The level of co-authorship of each published item and whether the co-authors worked in the same institution as the author studied, elsewhere in the UK, or abroad, were also recorded. The data were recorded in a spreadsheet, grouped by broad discipline categories.

2.5 Challenges

A number of challenges emerged during the data collection stages which have affected the analysis. The most significant of these are described below. In some cases, it might be possible to mitigate the effects by improving on the methods used; however this is not universal.

2.5.1 Identifying the correct researcher

Researchers were identified from their surname, initials, institutional affiliation and UoA. In the majority of cases this was sufficient, although there were occasional difficulties experienced in locating the correct individual. In particular, the 2003 sample was drawn from the RAE2001 list which was made up of all researchers returned from all categories, not just Category A (unlike the RAE2008 list which distinguishes between categories). Thus a small number of names were drawn in the sample with little or no continuing connection to UK higher education, making it difficult to ascertain whether the individual identified was the researcher we were looking for. However, this issue arose only in a small number of cases.

Another area of difficulty was in the arts subject area in both years, especially those names drawn from the Art and Design UoA. Here, most appeared to be practising artists, who had minimal presence on departmental websites, with the wider internet being only a little help.

2.5.2 Compiling accurate reference lists

There was no practical way of assessing the completeness of the lists of outputs obtained. Even where a researcher provided a detailed list of research outputs on their webpage, such lists could not be relied upon to be complete or up to date. Often, lists stopped at some point in the

recent past, notably the end of 2007 (at the end of the RAE2008 census period). Furthermore, adequate detail was not universal, with the majority of researchers sampled listing only selected outputs, or none at all, on their web pages.

Also, references on some individuals' lists were found to be incorrect in a number of cases. The most common errors concerned publications that were, presumably, forthcoming at the time of compiling the list but which were eventually published in a different year. This was mostly the case with outputs listed as being published in 2008, but which were subsequently found not to have been published in that year.

2.5.3 Researchers with no identifiable dissemination in the appropriate year

A large number of the sampled researchers showed no identifiable dissemination in the appropriate year, despite the universally acknowledged pressure to publish. This was the case even though an inclusive definition of dissemination was employed, following the lead of the RAE rules. Again, the arts showed a notable lack of visible output. Without doubt, in this case many, if not most, of the artists and musicians will have been producing work that was accessible by the wider world but this did not show up in the search results. Figure 3 gives details of the number and proportion of researchers for whom no outputs could be identified in 2003 and 2008.

Figure 3 Authors with no publications in the relevant years

	2003		2008	
	Number	% of sample	Number	% of sample
Bio-medicine	52	54%	44	46%
Physical sciences	23	32%	21	30%
Engineering	45	63%	26	37%
Social sciences	60	61%	44	46%
Humanities	24	30%	37	48%
Arts	38	73%	30	59%
Education	26	50%	29	57%
Total	269	52%	231	45%

2.5.4 Dissemination without citation

A number of items examined contained no citations, for example book reviews, conference outputs that were not published in formal proceedings or, in the case of music, CDs. This was more true of the Arts and Humanities than other disciplines, possibly because of a more informal approach to research gatherings, whether conferences, symposia etc.

2.5.5 Inability to locate an accessible copy of outputs

When attempting to retrieve non-WoS material, in some cases it proved impossible to obtain a copy at reasonable cost. Visits were made to a number of libraries, including the British Library at Boston Spa, but still some material could not be found, only being held overseas, at institutions that it was not possible to visit in the time and with the resources available, or simply because the item was on loan at the time of the visit.

2.5.6 Reference counting

Within the WoS cited reference lists, the references came in two forms, those indexed by WoS and those not indexed. The former were straight-forward and identification of the type of item being cited was simple. The latter required more skill, as these references contain both non-journal material and material that should have been indexed. WoS truncation of titles can often hide the identity of the cited work, but in the main an internet search provided enough information for the identification of problematic references.

In non-WoS reference lists, the categorisation of type was more subjective. One particular issue arose when one type of material had been (re)published as/in another. For example, journal articles or historical materials might be republished in books. In such cases, the form of the referenced material took precedence, rather than the substance. Other case by case decisions included whether an item was a book or grey literature, a book or an historical pamphlet, grey literature or website etc. Conference outputs were identified where possible, in whatever form they occurred, but not all such instances will have been possible to record. Websites in particular were problematic, given that much material was accessed via the WWW but was not a website, but another type of output.

2.5.7 Referencing systems

Three main styles of referencing system were employed by the majority of researchers: Harvard with an individual list of references (e.g. in articles or book chapters); Harvard with combined references (e.g. in edited books); and the Numerical system.

The Harvard system is straight-forward to count, and the whole article/chapter and bibliography were copied when a combined list of references was provided. The numerical system was most time-consuming to count, as references were often interspersed with other notes, but it was generally easy to avoid duplication owing to the convention of using a shortened title on repeated references. The most time consuming references were found in single author books where references were included in chapter notes at the back of the book, using shortened titles for repeats, but where the title of the work was stated in full in each subsequent chapter.

2.6 Notes on the analysis

As well as information on the material cited in each research output examined, contextual data were also collected covering the type of output, whether single or multiple-authored, and, if multiple, whether the authors were from the same or different institutions (and, if so, if this included non-UK institutions) and whether they were ordered alphabetically.

Three broad areas were investigated:

- What types and means of dissemination are used by researchers in different broad disciplines, and how has this changed over the last five years?
- To what extent do researchers in different broad disciplines collaborate with others when disseminating their work, and how has this changed over the last five years?
- How many items are authors citing in their work, and what types of output are these? How does this vary by discipline, and has it changed over time?

All analyses were carried out on weighted data, to draw inferences about practice across the UK as a whole. Weights were calculated based on the numbers of authors analysed in each

discipline compared to the distribution of academic staff in Figure 1. The object of this weighting was to remove the effects of the different balance of disciplines in the data file for each year, allowing comparisons between 2003 and 2008 overall. It has also ensured that the overall results provided valid estimates for UK HE as a whole. Wherever possible, formal analyses were also carried out between years within disciplines, and between disciplines within years.

Despite the deliberate bias in the sample towards Arts and Education, there were relatively few authors in the broad Arts discipline for whom research outputs could be identified in the two years analysed, or where copies of works could be obtained for examination. In around a dozen cases, these included performance-based outputs (e.g. exhibitions, music recordings) where we could assume no other material was formally cited. However, the number of authors included in the analysis from the Arts remained small. In order to make formal comparisons between the other, larger, disciplines included, Arts has been excluded from all analyses by discipline. The Arts, together with other disciplines for which data for any analysis were available from fewer than 20 authors, have not been shown separately in the tables and graphs in the description of the analyses and results; however the data have been included in all the totals presented.

Figure 4 gives details of the number of authors whose research outputs were examined and included in the analysis which follows. This is fewer than those with outputs identified as stated in Figure 2, owing to the difficulties in obtaining copies of some outputs for examination, described above. It also shows the weights applied to each discipline in the analysis. The total number of outputs analysed across both years was 1,452.

Figure 4 Calculation of weights

	2003			2008		
	No. analysed	Proportion of sample	Weight	No. analysed	Proportion of sample	Weight
Bio-medicine	44	19%	1.36	52	21%	1.22
Physical sciences	52	22%	0.60	47	19%	0.70
Engineering	27	11%	1.07	45	18%	0.68
Social sciences	37	16%	1.45	51	20%	1.12
Humanities	42	18%	0.77	20	8%	1.71
Arts	9	4%	1.74	15	6%	1.11
Education	24	10%	0.57	19	8%	0.76
	235	100%		249	100%	

The analysis was carried out using the SPSS® software package. Formal tests were carried out for statistical significance of apparent differences between years, between disciplines (excluding Arts), and in some analyses between types of source material, where there were sufficient data for the tests to be valid. All results reported as being statistically significant had a less than 5% probability of occurring by chance; the significance levels are noted in the text.

3 Results

3.1 Types of dissemination

Figure 5 shows the average number of outputs, of all types, found per author in each year, by discipline. There has been a small, but statistically significant, increase in the average from 2.5 in 2003 to 3.2 in 2008. Much of this is due to a considerable increase in Bio-medicine, which accounts for approximately one quarter of all UK HE researchers (*Figure 1*).

Figure 5 Average number of outputs

	2003		2008		
	Mean	Standard error	Mean	Standard error	Significant?
Bio-medicine	2.32	.188	4.52	.597	p<0.01
Physical sciences	4.29	.851	4.51	.609	no
Engineering	3.22	.659	3.31	.474	no
Social sciences	2.00	.198	2.69	.266	p<0.05
Humanities	1.93	.303	1.70	.124	no
Education	2.17	.631	1.88	.245	no
Total, inc. Arts	2.50	.170	3.19	.206	p<0.05

Figure 6 illustrates the proportions of outputs by type for each discipline. There were insufficient data to test the apparent differences by discipline, either overall or within year. Combining disciplines, the difference between years overall was found to be statistically significant (p<0.01). There is a greater proportion of journal articles in 2008, and of editorial material, meeting abstracts, and 'other' types of material; conversely there are lower proportions of books, book chapters, conference proceedings and book reviews. (Note that the data have been weighted to reflect the population distribution of disciplines, so that changes in the disciplinary distribution do not account for the difference over time.)

Figure 6 Outputs by type

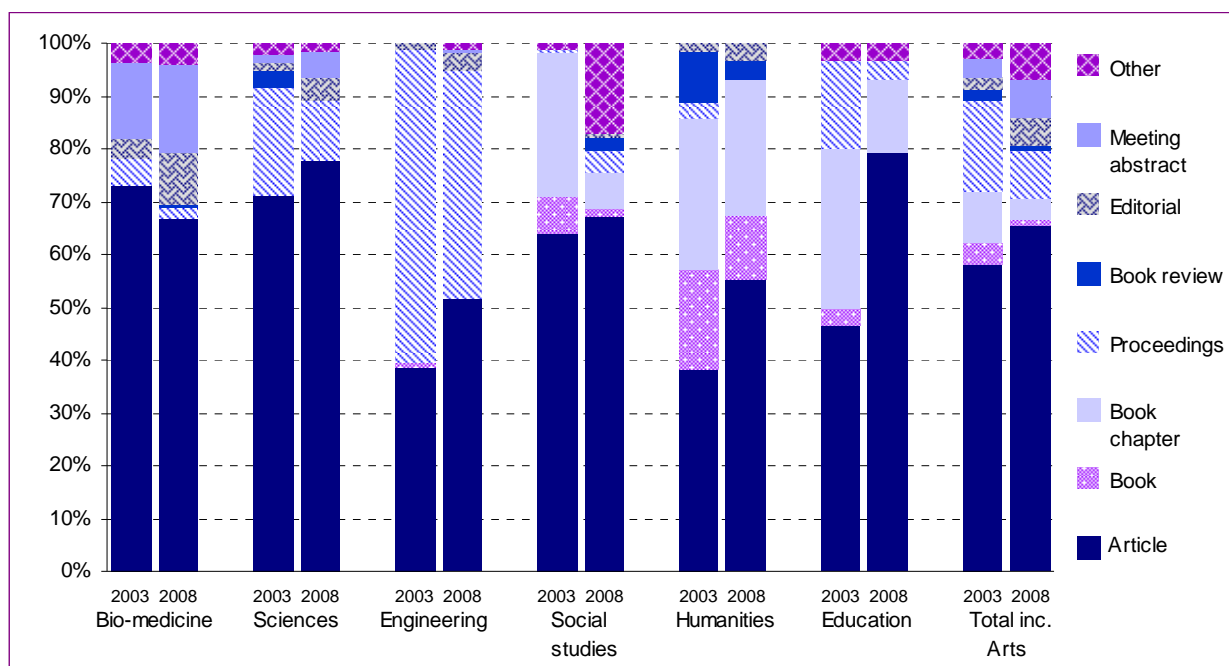
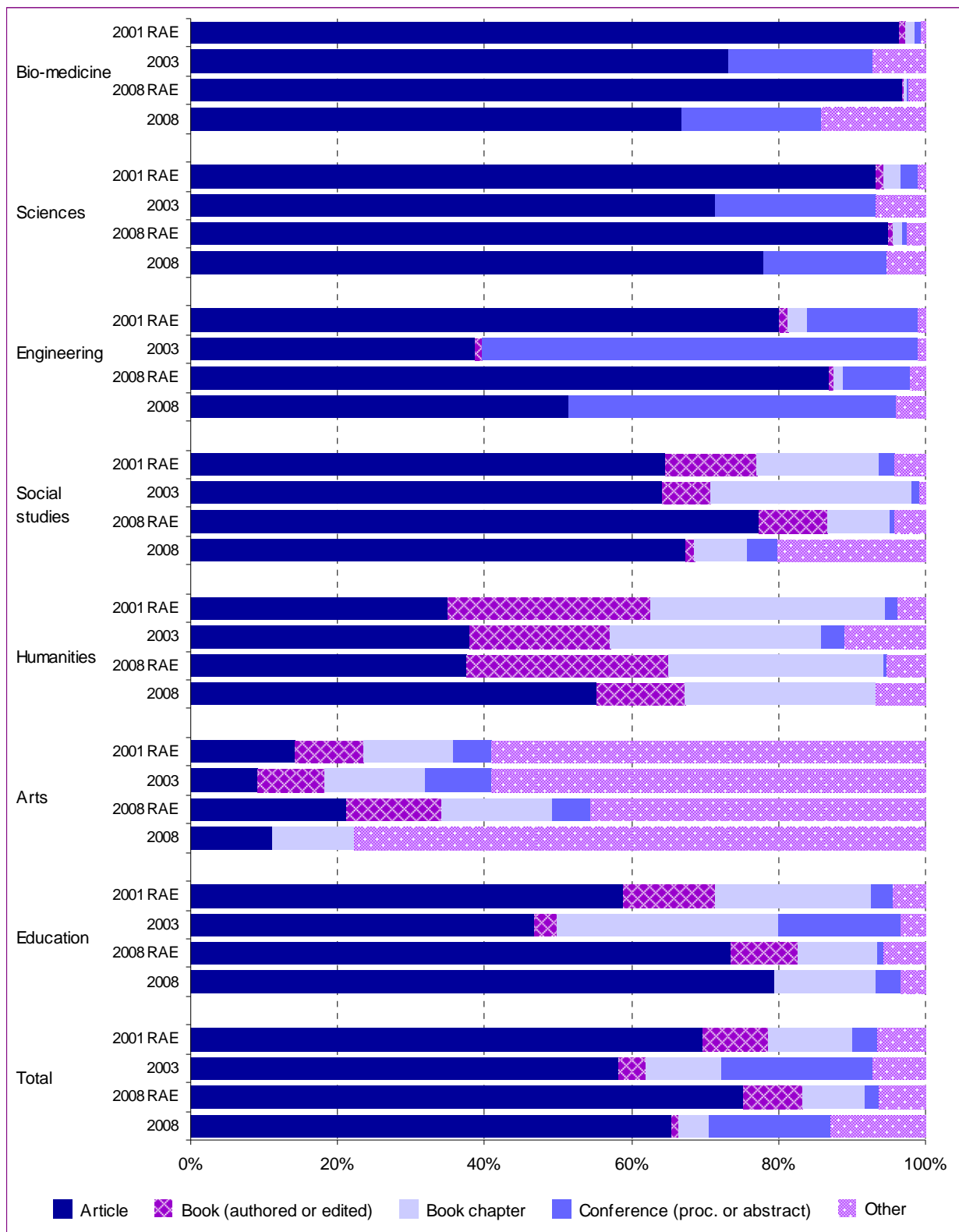


Figure 7 Comparison of outputs produced with those submitted to RAE



It is of interest to compare this distribution with that of the RAE submissions for 2001 and 2008. Summarising output types into five broad categories, Figure 7 shows that, for all disciplinary groups except Humanities, there was greater emphasis on journal articles in the outputs submitted to the RAE than in the publication patterns revealed by the bibliometric analysis. This

indicates that the outputs submitted to the RAE do not always reflect the actual disciplinary patterns for research dissemination.

Engineering shows a much greater use of conference proceedings in its output as compared to its RAE submissions, although it is unclear whether there is replication of conference material in the submitted articles. The Humanities shows a smaller proportion of journal articles submitted to the RAE than was identified in the bibliometric study, most notably in 2008. The Arts show the highest proportion of ‘other’ outputs, both in the RAE submissions and in the small number of outputs found in the bibliometric study. Here, the creative nature of the constituent subjects is evident in the types of material encompassed by ‘other’: artefact, composition, exhibition, performance, scholarly edition etc. This is evidence of the inclusive nature of the RAE criteria in terms of outputs that were eligible to be assessed.

3.2 Collaboration

Figure 8 shows the proportions of single and multiple authored works by discipline. Overall, there is a statistically significant increase in the proportion of multiple author works between 2003 and 2008 ($p < 0.01$). On average, 86% of works were recorded as having multiple authors in 2008, compared to 76% in 2003. Excluding Arts from the analysis (insufficient data for comparison), and taking both years together, there is also a statistically significant difference ($p < 0.01$) between disciplines in the proportions of multiple authored works, clearly illustrated in Figure 8. Considering each discipline individually, there are statistically significant differences between years for Physical sciences (from 90% in 2003 to 96% in 2008 - $p < 0.05$) and Social sciences (from 57% in 2003 to 84% in 2008 - $p < 0.01$). Humanities show over two-thirds of the analysed works as being by single author works, a much higher proportion than the other disciplines.

Figure 8 Numbers of authors per output

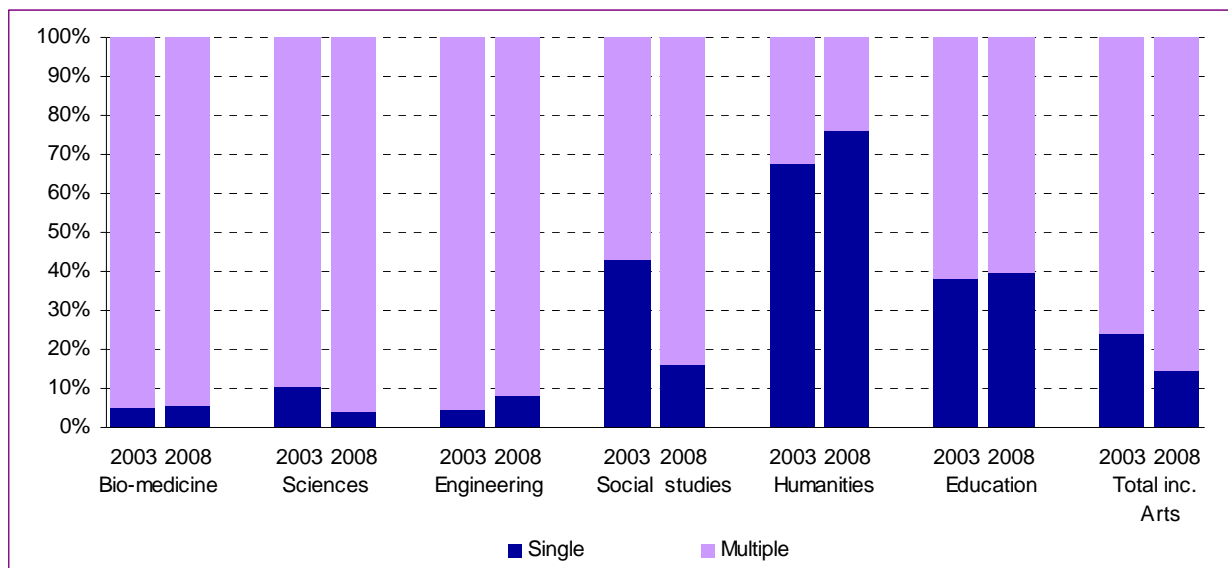
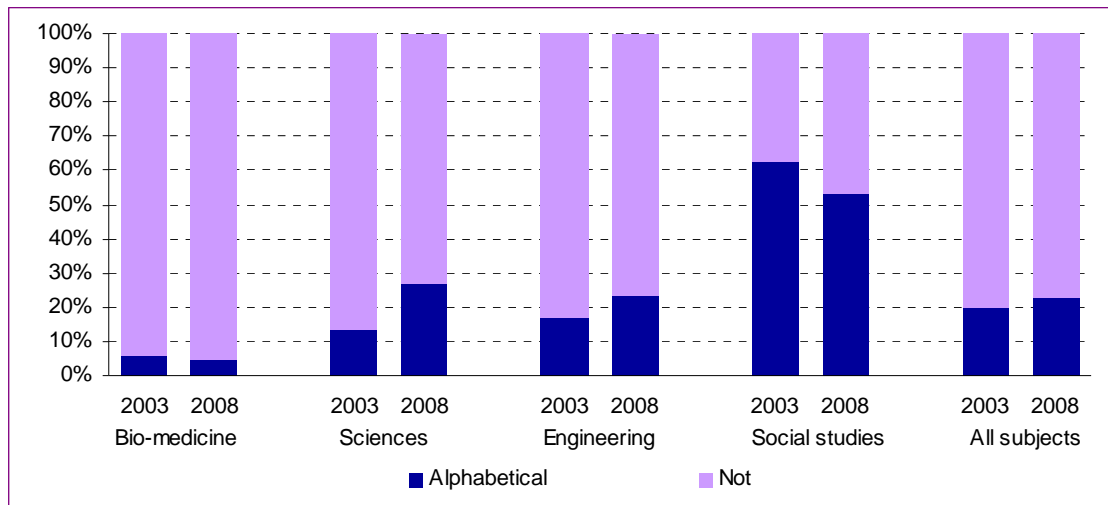


Figure 9 is derived from those works which had more than one author, and shows the proportions of these which listed their authors in alphabetical order.

Figure 9 Authorship order



Overall, there is no difference between years in the proportions of outputs where authors are listed alphabetically. Excluding Arts from the analysis (insufficient data for comparison), and taking both years together, there is a statistically significant difference ($p < 0.01$) between disciplines in the proportions listing authors alphabetically, illustrated in Figure 9. Looking at each discipline individually, there is a statistically significant difference for sciences ($p < 0.01$), where the proportion of works listing authors alphabetically increased from 13% in 2003 to 27% in 2008; and insufficient data to test in Humanities and Arts.

Figure 10 is also derived from those works which had more than one author, and shows the proportions of these where authors were associated with two, or more, institutions. There is a statistically significant increase ($p < 0.01$) in the proportion of works with inter-institutional collaborations overall, from 62% in 2003 to 73% in 2008. Excluding Arts from the analysis (insufficient data for comparison), and taking both years together, there is also a statistically significant difference ($p < 0.01$) between disciplines in the proportions of works with inter-institutional collaborations, illustrated for the larger disciplines in Figure 10. Looking at the various disciplines individually, there are statistically significant differences over time for Bio-medicine, for which the proportion of multiple institution collaborations has increased from 74% to 82% ($p < 0.05$), and for Physical sciences, which has increased from 63% to 78% ($p < 0.01$). There is insufficient data to test Humanities, Arts and Education.

Figure 10 Institutional collaborations

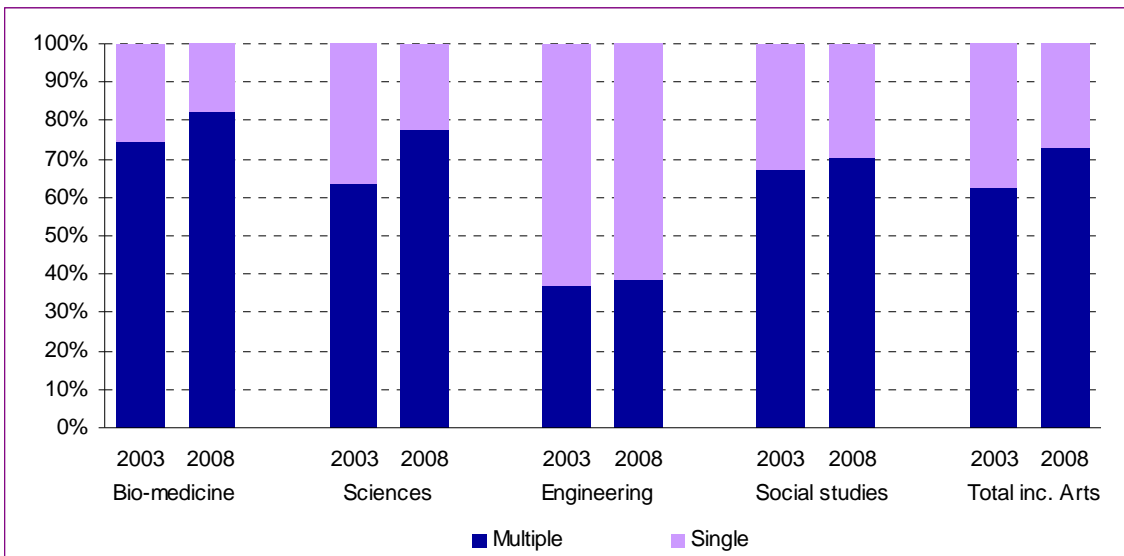
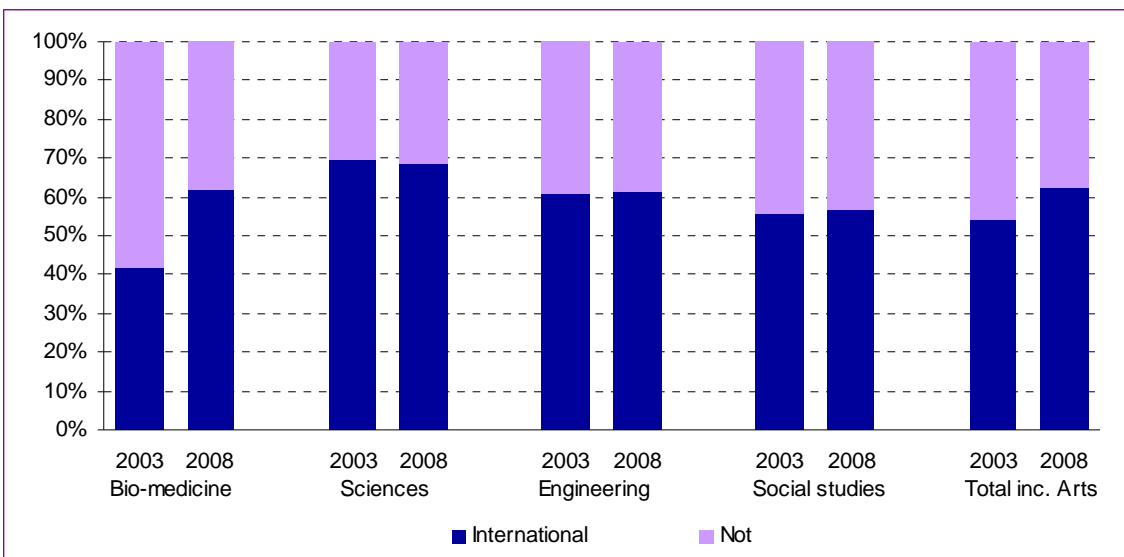


Figure 11 is derived from those works which had authors from more than one institution, and shows the proportions of these with international collaborations. Overall there is a statistically significant ($p < 0.05$) increase in the proportion of works involving international collaboration, from 54% in 2003 compared to 62% in 2008. Excluding Arts from the analysis (insufficient data for comparison), and taking both years together, there is also a statistically significant difference ($p < 0.05$) between disciplines in the proportions of works with international collaborations, illustrated in Figure 11 for those disciplines where the numbers are based on 20 or more authors. Looking at the various disciplines individually, there is a statistically significant difference only for Bio-medicine ($p < 0.01$), which increased from 42% in 2003 to 62% in 2008. There was insufficient data to test apparent differences in Humanities, Arts and Education.

Figure 11 International collaborations



3.3 Citation practice

Figures 12 and 13 summarise the data gathered on citation practices. Overall, 6.8% of the textual works examined did not include any citations in 2003, compared to 10.8% in 2008. This difference overall was found to be statistically significant ($p < 0.05$). Figure 12 illustrates this by discipline; apparent differences between years for individual disciplines were found not to be statistically significant where there was sufficient data to carry out the test. The inclusion or not of citations also tends to be associated with the type of output, with meeting abstracts, editorial pieces and book reviews least likely to include citations (Figure 13), although there was insufficient data to formally test the apparent differences.

Figure 12 Inclusion of citations by discipline and year

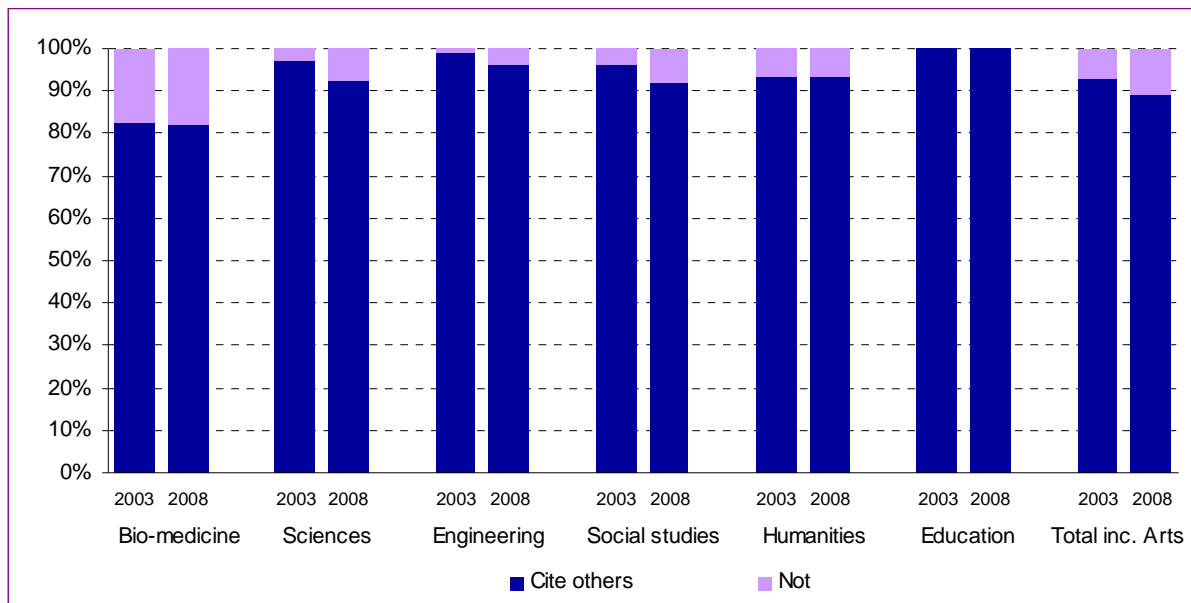


Figure 13 Inclusion of citations by source

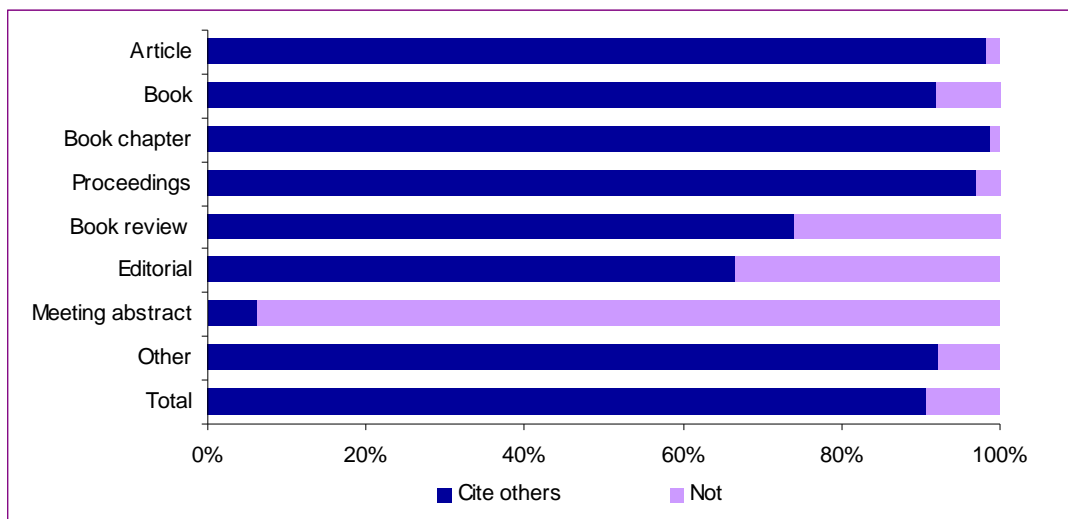


Figure 14 shows the average number of citations per output, both by discipline and by output type. Overall, the difference in the average number of citations per output between 2003 and 2008 was not found to be statistically significant. There are, however, differences in the average

number of citations per output between types of output and, consequentially, between disciplines.

Figure 14 Average number of citations per output

	Mean	Standard error	Significant?
2003	37.07	2.600	
2008	33.68	1.141	no
Bio-medicine	28.51	1.259	
Physical sciences	29.68	1.498	
Engineering	18.99	.963	
Social sciences	53.13	3.246	
Humanities	59.75	10.929	
Arts	39.36	14.337	
Education	36.67	5.624	p<0.01
Article	37.55	.852	
Book	230.96	42.537	
Book chapter	47.16	4.560	
Proceedings	19.48	1.575	
Book review	3.95	1.546	
Editorial	8.90	1.731	
Meeting abstract	1.16	.706	
Other	28.52	5.268	p<0.01
Overall	35.14	1.293	

Constructing a model including both source and discipline, analysis of variance indicates that, as well as the statistically significant differences in mean numbers of citations per output by discipline and by source type illustrated in Figure 14, there is also a statistically significant interaction between these factors. This suggests that the average number of citations per output for the various types of source outputs differ according to discipline. However, this result should be treated with caution, as there are several discipline/source type combinations which were not observed in the data collected. It seems likely that one potential explanation for the interaction effect observed may be that the data collection method resulted in the identification of almost no books as published research outputs in the Bio-medicine, Science and Engineering disciplines.

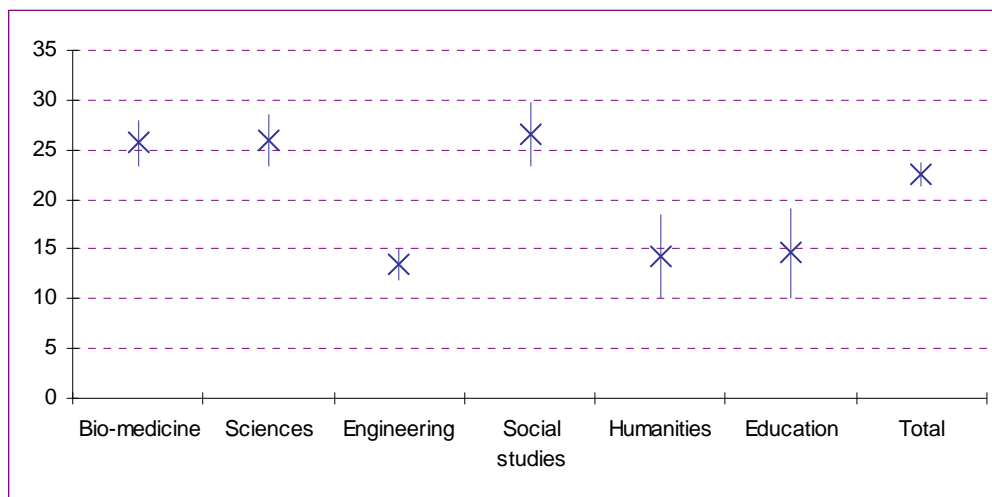
Although there is no difference over time in the total number of citations per output, there are statistically significant differences in the types of work being cited, shown in Figure 15. There are more citations to journal articles and websites in 2008 than in 2003, but fewer citations to books and grey literature. There are also differences in the average numbers of citations of different formats between disciplines ($p<0.01$ for all types of cited material), and between types of source output ($p<0.01$ for all types of cited material except outputs in press, where $p<0.05$).

Figure 15 Average number of citations per output, by type of material cited

	2003		2008		Significant?
	Mean	Std. Error	Mean	Std. Error	
Articles	20.0	0.94	24.3	0.83	p<0.01
Books	11.6	1.89	5.5	0.53	p<0.01
Conference outputs	0.9	0.15	0.8	0.09	no
Grey literature	2.1	0.35	1.2	0.12	p<0.05
Websites	0.3	0.06	0.3	0.10	no
Theses	0.2	0.03	0.2	0.02	no
In press	0.2	0.02	0.1	0.02	no
Other	2.0	0.38	1.2	0.17	p<0.05
Total	37.1	2.60	33.7	1.14	no

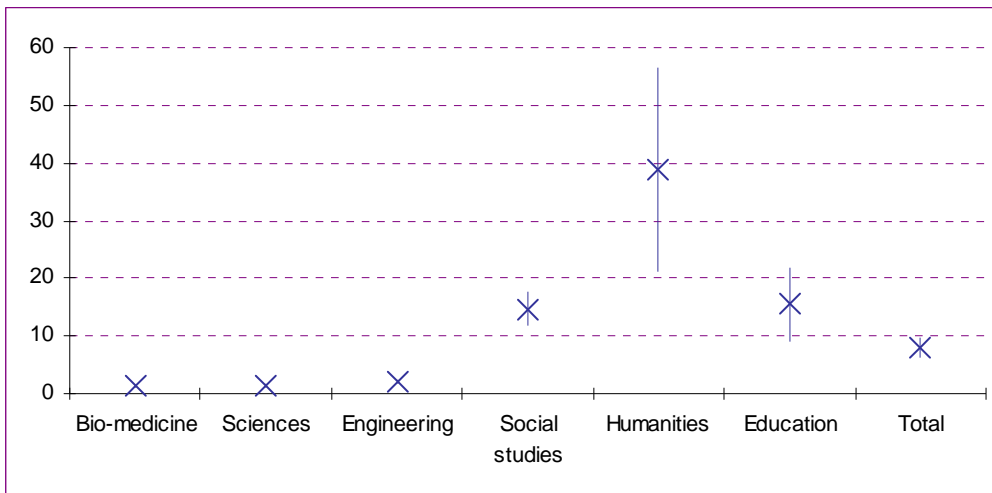
Figure 16 to Figure 20 illustrate the mean number of citations per output by discipline, for selected types of material cited. In all cases, the differences between subjects were found to be statistically significant ($p<0.01$). While all disciplines cite journal articles, Bio-medicine, Physical sciences and Social sciences cite around twice as many, on average, as Engineering, Humanities and Education (Figure 15). Humanities cite the most books, on average, with Social sciences and Education also citing these to some extent (Figure 17). Engineering cites more conference outputs than do Bio-medicine, Physical sciences or Social sciences (Figure 18). Social sciences cite the most grey literature, on average, with Education also citing these, and to a lesser extent, Humanities (Figure 19). Web sites are rarely cited, but Social sciences, Education and Humanities are more likely to do this than the other disciplines (Figure 20) (however, it is clear from the data collection process that internet sources for other types of material are used more often).

Figure 16 Average number of journal article citations per output



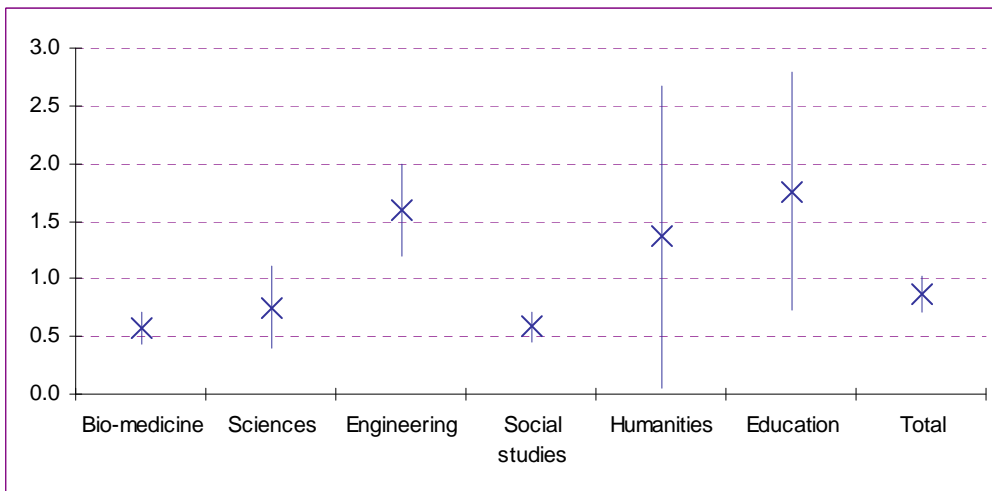
X is the mean; the bars indicate the standard error of the mean

Figure 17 Average number of book citations per output



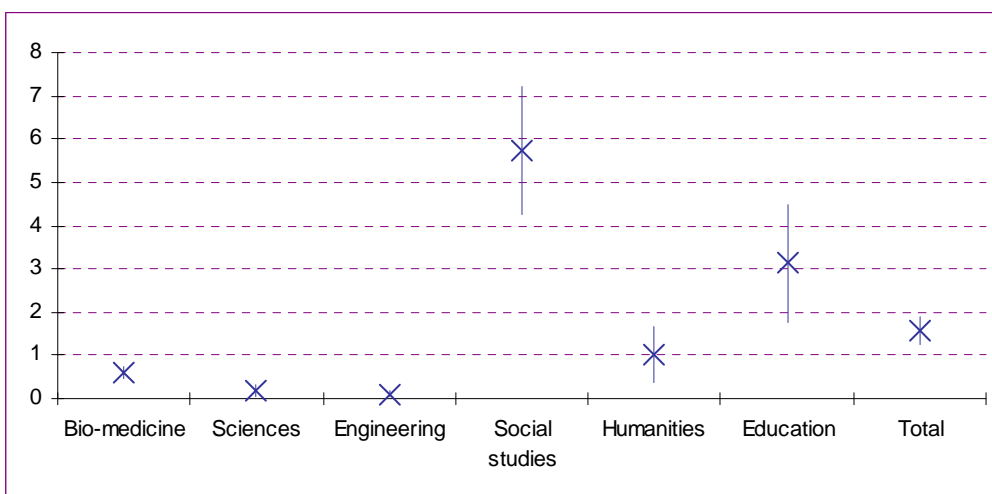
X is the mean; the bars indicate the standard error of the mean

Figure 18 Average number of conference outputs cited per output



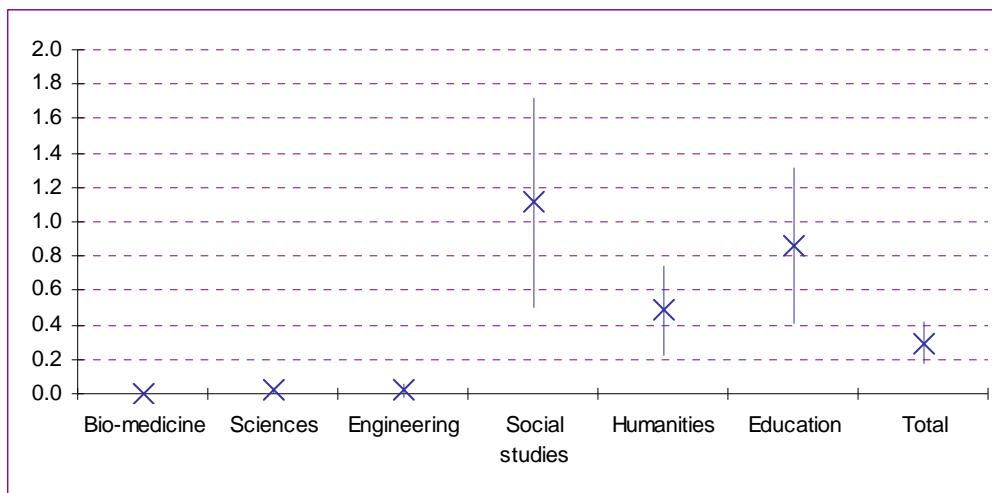
X is the mean; the bars indicate the standard error of the mean

Figure 19 Average number of grey literature items cited per output



X is the mean; the bars indicate the standard error of the mean

Figure 20 Average number of web sites cited per output



X is the mean; the bars indicate the standard error of the mean

Figure 21 shows the proportions of material of different types cited in a selection of output types. While journal articles are the most frequently cited works overall, books and book chapters are most likely to cite books/book chapters. Conference outputs are most likely to be cited in other conference proceedings

Figure 21 Proportion of citations, by type of material cited and output type

Citing work	Type of cited material								Total
	Articles	Books*	Conf'ce	Grey lit	Other	URLs	Theses	In press	
Article	74%	15%	2%	3%	4%	1%	0%	0%	100%
Book	27%	64%	1%	5%	3%	0%	1%	0%	100%
Book chapter	31%	47%	2%	8%	9%	1%	1%	0%	100%
Proceedings	71%	10%	8%	2%	7%	0%	1%	1%	100%
Editorial	84%	11%	1%	2%	1%	0%	0%	1%	100%
All outputs	64%	23%	2%	4%	4%	1%	1%	0%	100%

* Includes book chapters

4 Discussion and conclusions

Given the breadth of this bibliometric study, designed to cover the full range of research in the UK, the data collection was an extensive undertaking. There was no way to automate the process and, as such, many researcher-hours were spent in collecting the references of over 1,400 works, followed by considerable amounts of analysis of reference types and the subsequent counting of these. Consequently, it is understandable why previous studies of citation behaviour have concentrated on detailed analysis within small areas of research or broad analysis of wider disciplines, rather than showing a detailed analysis of a comprehensive range of research as presented here. The time required to undertake such exercises should not be underestimated.

One important finding from this investigation was the number of authors in both samples which did not appear to have any identifiable publications in the years being analysed. By using the 2001 and 2008 RAE submission lists as the sampling frame, the population of potential authors from which the samples were drawn might all be defined as 'research-active', and although some will inevitably be more active than others, around half of those in the sample did not have published research which we could identify, across all subject areas. This is a relatively high proportion, given the distribution of the number of publications amongst those authors for whom material was identified, and further investigations in this area would be of interest.

The data in Figure 7 show that there are differences between the research outputs which are submitted for research assessment purposes and scholarly activity as represented by the present sample of works. Perhaps the most notable difference is that of the proportions of journal articles and conference proceedings in Biomedicine, Science and Engineering, where journal articles form a much greater part of RAE submissions, with conference proceedings rarely being submitted. Indeed, the importance of journal articles within the RAE is demonstrated by the fact that the proportion of submitted outputs that are journal articles increased between 2001 and 2008 in each of the seven subject areas.

The disparity between RAE submissions and the picture of UK research as found in this study could be seen to confirm evidence of the highly selective nature of output submission in the RAE, as shown in other parts of this research. A partial explanation of the differences between the proportions of output types in the bibliometric analysis and the RAE may be that the bibliometric analysis found material that was unlikely to have been submitted to RAE. This would include more peripheral work such as posters and editorial material, as well as that which might be superseded by later work. For example, a conference abstract may become a full conference paper or journal article, and all formats, if found in the relevant year, would have been included in the bibliometric analysis, while only the more substantive work would have been considered for the RAE.

Looking at the data in further detail, the analysis found an increase in the average number of outputs per author overall, particularly in Biomedicine and Social sciences, but reasons for this are not clear. One suggestion is that more shorter works may have been produced, for example journal articles, book reviews and conference abstracts and fewer longer works, such as monographs, book chapters and proceedings papers. There are insufficient data available in this analysis to test this hypothesis in detail by subject. There are, however, some indications that there were fewer journal articles and more abstracts and editorial material produced in

Biomedicine in 2008 than in 2003. Restrictions on word counts reported in other strands of the research may also lead authors to prepare two shorter outputs where previously they may only have written a single, longer, one.

An alternative reason could be the increase in collaborative research and multi-authored papers which was found overall, and in Physical sciences and Social sciences, but not in Biomedicine, where the proportion of multiple-authored works is high in both years. Data were not collected regarding the number of authors for each output, which might have shed further light on this aspect. Note that there was no overall increase in the average number of outputs per author in Physical sciences, however.

There was a greater proportion of outputs which did not cite any references in 2008 than in 2003. This is likely to be associated with the types of output, and there are indications that there were more editorial pieces, abstracts and book reviews in 2008 than in 2003. There were insufficient data to test this formally, however.

Although there was no statistically significant difference in the average number of citations per output between the two years examined, the observed mean for 2008 was, in fact, lower than in 2003. Other areas of this research have found evidence of journal practices limiting the numbers of references allowed.

There were some interesting indications of interactions between subject and publication type when considering citation behaviour. It is only to be expected that monographs will include more references than journal articles, but there are some indications that there are disciplinary differences, particularly in the numbers of references included in monographs. This is an area where further investigations would be interesting.

This study attempted to count websites as a separate category. As noted above it was found to be challenging to distinguish between websites as an output in their own right and websites as a vehicle for dissemination. It was clear from the data collection process that the internet has become a much more widely used route to research outputs, with many more instances of such use being seen in 2008. However, no data were collected to reflect this. As it stands, the Social sciences do show a greater use of websites in research, though the very small proportion of these compared to other outputs shows that they continue to be a relatively minor area of use. Biomedicine, Science and Engineering show almost no use of websites but this is probably due to the abbreviation of such references by WoS making them very hard to detect.

4.1 Recommendations concerning methodology

If RAE lists, or equivalent, are used to sample academic staff, a larger initial sample is required to allow for individuals with no outputs in the year of interest. An additional bias towards the Arts is also indicated, to allow this area to be included in valid comparisons of practices across disciplines.

Additional value would be added by counting the number of authors, and of institutions, in multiple author works. Although we have been able to show an increase in multiple authored works, we have not been able to investigate whether there has been an increase in the number of authors, as suggested by the literature review.

The method for obtaining lists of research outputs in the biomedical, physical science and engineering disciplines was biased towards journal articles, and uncovered only one book and no book chapters. The survey showed that monographs in particular are not considered particularly important in these disciplines, but they are used to maximise dissemination to particular target audiences. Although using WoS was an efficient data source, it should be supplemented by other methods to discover a wider variety of outputs in these disciplines.

If repeated, it would be instructive to revise the method regarding the classification of websites by including a non-exclusive category for references with URLs to in order to better reflect the use of the internet as an information source.

Appendix: Allocation of subjects to broad disciplines

Medical sciences

RAE 2001

Biological Sciences; Clinical Laboratory Sciences; Community-based Clinical Subjects; Hospital-based Clinical Subjects; Clinical Dentistry; Pre Clinical Studies; Anatomy; Physiology; Pharmacology; Pharmacy; Nursing; Other Studies and Professions Allied to Medicine; Psychology

RAE 2008

Biological Sciences; Cardiovascular Medicine; Cancer Studies; Infection and Immunology; Other Hospital Based Clinical Subjects; Other Laboratory Based Clinical Subjects; Epidemiology and Public Health; Health Services Research; Primary Care and Other Community Based Clinical Subjects; Psychiatry, Neuroscience and Clinical Psychology; Dentistry; Nursing and Midwifery; Allied Health Professions and Studies; Pharmacy

HESA Academic cost centres

Biosciences; Clinical medicine; Clinical dentistry; Anatomy & physiology; Nursing & paramedical studies; Health & community studies; Psychology & behavioural sciences; Pharmacy & pharmacology

Physical sciences

RAE 2001

Agriculture; Food Science and Technology; Veterinary Science; Chemistry; Physics; Earth Sciences; Environmental Sciences; Library and Information Management; Pure Mathematics; Applied Mathematics; Statistics and Operational Research

RAE 2008

Pre-clinical and Human Biological Sciences; Agriculture, Veterinary and Food Science; Chemistry; Physics; Earth Systems and Environmental Sciences; Library and Information Management; Pure Mathematics; Applied Mathematics; Statistics and Operational Research

HESA Academic cost centres

Chemistry; Physics; Earth, marine & environmental sciences; Mathematics; Veterinary science; Agriculture & forestry

Engineering

RAE 2001

General Engineering; Chemical Engineering; Civil Engineering; Electrical and Electronic Engineering; Mechanical, Aeronautical and Manufacturing Engineering; Mineral and Mining Engineering; Metallurgy and Materials; Computer Science

RAE 2008

Electrical and Electronic Engineering; General Engineering and Mineral & Mining Engineering; Chemical Engineering; Civil Engineering; Mechanical, Aeronautical and Manufacturing Engineering; Metallurgy and Materials; Computer Science and Informatics

HESA Academic cost centres

General engineering; Chemical engineering; Mineral, metallurgy & materials engineering; Civil engineering; Electrical, electronic & computer engineering; Mechanical, aero & production engineering; Information technology & systems sciences & computer software engineering

Social sciences

RAE 2001

Built Environment; Town and Country Planning; Geography; Law; Anthropology; Economics and Econometrics; Politics and International Studies; Social Policy and Administration; Social Work; Sociology; Business and Management Studies; Accounting and Finance

RAE 2008

Architecture and the Built Environment; Town and Country Planning; Geography and Environmental Studies; Law; Anthropology; Economics and Econometrics; Politics and International Studies; Development Studies; Social Work and Social Policy & Administration; Sociology; Business and Management Studies; Accounting and Finance; Psychology

HESA Academic cost centres

Architecture, built environment & planning; Catering & hospitality management; Business & management studies; Geography; Social studies; Media studies

Humanities

RAE 2001

American Studies; Middle Eastern and African Studies; Asian Studies; European Studies; Celtic Studies; English Language and Literature; French; German, Dutch and Scandinavian Languages; Italian; Russian, Slavonic and East European Languages; Iberian and Latin American Languages; Linguistics; Classics, Ancient History, Byzantine and Modern Greek Studies; Archaeology; History; History of Art, Architecture and Design; Philosophy; Theology, Divinity and Religious Studies

RAE 2008

American Studies and Anglophone Area Studies; Middle Eastern and African Studies; Asian Studies; European Studies; Celtic Studies; English Language and Literature; French; German, Dutch and Scandinavian Languages; Italian; Russian, Slavonic and East European Languages; Iberian and Latin American Languages; Linguistics; Classics, Ancient History, Byzantine and Modern Greek Studies; Archaeology; History; History of Art, Architecture and Design; Philosophy; Theology, Divinity and Religious Studies

HESA Academic cost centres

Humanities & language based studies; Archaeology; Modern languages

Arts

RAE 2001, 2008

Art and Design; Communication, Cultural and Media Studies; Drama, Dance and Performing Arts; Music

HESA Academic cost centres

Design & creative arts

Education

RAE 2001, 2008

Education; Sports-related Subjects

HESA Academic cost centres

Education; Sports science & leisure studies; Continuing education

About the Research Information Network

Who we are

The Research Information Network has been established by the higher education funding councils, the research councils, and the national libraries in the UK. We investigate how efficient and effective the information services provided for the UK research community are, how they are changing, and how they might be improved for the future. We help to ensure that researchers in the UK benefit from world-leading information services, so that they can sustain their position as among the most successful and productive researchers in the world.

What we work on

We provide policy, guidance and support, focusing on the current environment in information research and looking at future trends. Our work focuses on five key themes: **search and discovery, access and use of information services, scholarly communications, digital content and e-research, collaborative collection management and storage.**

How we communicate

As an independent voice, we can create debates that lead to real change. We use our reports and other publications, events and workshops, blogs, networks and the media to communicate our ideas. All our **publications** are available on our website at www.rin.ac.uk

The full report is available at www.rin.ac.uk/communicating-knowledge, along with these supporting papers. Hard copies can be ordered via email contact@rin.ac.uk

Get in touch with us

The Research Information Network
96 Euston Road
London
NW1 2DB
UK

Telephone +44 (0)20 7412 7946

Fax +44 (0)20 7412 7339

Email contact@rin.ac.uk