Doing data analysis

Stephen Gorard

Introduction: four kinds of evidence

Research is about more than empirical evidence, but evidence is at the heart of finding out more about the social and education world. One way of marshalling evidence on a topic, or to answer a research question, is to use the findings of others as published in the literature. This use of evidence at third-hand is common – in the notorious literature review for a PhD, for example. I say 'third-hand' because the analyst does not have access to the primary evidence, nor are they re-presenting an analysis of the data. They are presenting a summary of what a previous author presented about an analysis of data. Done well, with a clear focus, such a review of literature can be useful, at least in establishing what others think, how a topic is usually researched, and why the topic might be important to research further. Some of the inherent weaknesses of using the accounts of others might be overcome by ensuring that all of the relevant literature was used, even accounts of unsuccessful studies and evidence from unpublished studies, and then conducting a full meta-analysis of the results (I recommend using a Bayesian approach, see appendix to Gorard et al. 2004, which allows the relatively simple combination of different kinds of evidence). But such systematic reviews of evidence are rare, very difficult to do properly, and both expensive and time-consuming. And anyway this second approach does not overcome the chief drawbacks of the literature which are that we have no direct access to the evidence of others, and often face a very partial view of the assumptions made and the analyses conducted.

Much better, in many ways, is the approach of collecting primary data yourself and conducting the cleaning, coding and analysis yourself. This third kind of evidence overcomes many of the drawbacks you will encounter in trying to understand what other people have done from their own accounts (but try to remember to make your own accounts suitably clear about the assumptions, short cuts, and compromises you have made). It is probably true to say that most education researchers use primary evidence at some stage, just as most of them conduct reviews of literature. The drawbacks of generating primary evidence include the time and cost involved, and so the likely small scale of your own study.

A compromise, used by only a minority of social science researchers at present, lies in the re-analysis of secondary data. Secondary data has been generated by others but is available directly to new researchers to conduct their own analyses. Many of the techniques, craft tips, and issues covered in this chapter may have applications beyond the use of secondary data, but secondary data is the focus of what follows. I consider in turn some of the likely sources of this fourth kind of evidence, why all of us should perhaps be using secondary evidence more, and how such data might be analysed.

Where do we find secondary data?

Much of the existing data that might be useful for education and social science researchers is available for download from websites, or can be requested from official bodies via their websites. I suggest some likely sources here for illustration, predominantly from the UK, but the details of internet resources are likely to date rapidly, and to vary between countries.

The (Office for) National Statistics is a one-stop shop for evidence on almost anything. It includes evidence at small area level on all ten-yearly national censuses of the population, most recently from 2001, and next run in 2011 (http://www.statistics.gov.uk/census2001/topics.asp). Here you can find such things as the highest educational qualification of everyone in the population aged 16-74, broken down by sex, age, area of residence, type of accommodation, health, religion, occupation, marital status, and so on. You can also request bespoke tables and specific analyses.

The national UK Data Archive (http://www.data-archive.ac.uk/) is a repository of all datasets generated through research paid for by the taxpayer-funded Research Councils (such as the Economic and Social Research Council), and from a number of other sources. It includes historical archives, policy and other documents, and transcripts of interviews undertaken as part of previous research projects. Some of it is relevant to education studies. You can register for access to these resources, and then reanalyse the evidence for your own purposes.

The National Digital Archive of Datasets (http://www.ndad.nationalarchives.gov.uk/) similarly contains a wide range of data – including a database of the annual schools census for all schools in England, undertaken in January each year, collecting data at school level on pupil intake characteristics (poverty, special needs, ethnicity, sex, first language) and on the teaching and support staff. The Department for Children, Schools and Families has a website full of data on all aspects of school and childhood, including an archive of examination and key stage results for each school up to the current year (http://www.dcsf.gov.uk/performancetables/). Linking school and pupil characteristic data from the annual census to the corresponding records of school examination entry and attainment is a common but influential approach to secondary data analysis.

Edubase is a set of data about every educational institution in England, summarising their intake, management, whether they are in special measures (for schools) and even including information on the population density of the locale. The publicly available component can be accessed at http://www.edubase.gov.uk/home.xhtml. More complex but even more detailed are two related databases held by central government and made available to researchers on request. These databases are not more generally available since they might identify individual students, and researchers can only request them for a specific study that meets ethical and data protection approval, and they must return or destroy the figures after completion. The pupil level annual schools census (PLASC, also now increasingly just called ASC) contains a record for every pupil in maintained schools in England. It details their background characteristics, including periods in-care, special needs status and first language. It also has some attainment data. The national pupil database (NPD) holds individual records on every pupil in maintained schools in England. It details their examination and assessment entry and attainment, and also has some background data. An

application for both datasets would be via the DCSF. There are some equivalent datasets for other home countries. See, for example, http://www.npd-wales.gov.uk/.

The PLASC/NPD combination, perhaps with Edubase as well, is very powerful. The individual records for schools and pupils can be matched across datasets. You could find out whether pupils eligible for free school meals in villages enter more GCSEs than equivalent pupils in cities – if you wanted to. Even more importantly, the records can be matched across years, so that individual pupils can be tracked annually from the moment they enter the system. This means that you could relate subsequent patterns of post-16 participation for each ethnic group to their earliest primary school qualification, for instance. These datasets are widely used in policy and practice, perhaps most notably in calculation of contextualised value-added scores which are meant to provide measures of school standards and effectiveness, and so inform a range of processes from OFSTED inspections to parental choice (Gorard 2006a, 2008a, 2010a).

Many other official and government websites include downloadable data. One of these is the Higher Education Statistics Agency (http://www.hesa.ac.uk/). Here you can find an archive of applications and admissions to higher education, and discover changes over time or regional variations in what kind of students study what kinds of subjects at university, for example. Or what happens to patterns of application for HE following a new policy, such as the introduction of student loans (Gorard et al. 2007).

Beyond the UK, the OECD website has a collection of international educational evidence, including the annual Education at a Glance which has sections on work-based and tertiary education as well as schooling. It also contains the results of successive rounds of the international PISA study. The most recent PISA study at time of writing was in 2006 (http://pisa2006.acer.edu.au/), and the database includes the views of teachers and students, student test results in a range of subjects, and school-level data. It can be downloaded from the website, giving records for individuals within schools, in around 80 countries. If you want comparative data about schools, teachers or education systems, or to place your evidence in an international context, then there is unlikely to be a better source of ready-made evidence. An example of re-use of PISA data appears in Gorard and Smith (2004).

For more on where to find data, and how to analyse it, see Gorard (2003) and Smith (2008). The examples here only touch the surface of the local, national and international datasets made available specifically so that you and I can use them for our own purposes. Whatever you want to know about education it is very likely that someone has already collected the evidence you need on a larger scale and with higher quality than your resources would allow. Perhaps the most original new use of these existing datasets lies in combining evidence from two or more in a way that has not been done before. Can you imagine linking the schools data with the HESA data – who is qualified for but missing out on university? You could use the resident population census data with the schools intake data. Do the pupils in different kinds of schools represent their local residents, or do faith-based schools ‘select’ by socio-economic background as well as religion? You might try to decide whether pupil views on citizenship (from PISA 2006) are related to the type of school they attend (annual schools census). One of the many interesting projects I conducted involved comparing present day stories of adult learning with those in the taped oral archive of

families living in the South Wales coalfields in the 1890s (Gorard and Rees 2002). There are many such possibilities, but they are often ignored by scholars.

Why might we use secondary data?

Where you are able to gain direct access to the evidence collected by others this allows you a larger scale and range of data than you could collect yourself, but still with many of the advantages of primary data in terms of knowledge and control. A range of existing evidence is available on almost all social science topics. Such secondary evidence has the disadvantages, for you, that it was usually collected for another purpose and so may not be ideal, that you may have little idea of the conditions under which the data was collected, and you may therefore be misled about its completeness and accuracy. Nevertheless, when considering an education issue, secondary evidence is usually about as useful as primary evidence, immediately available, larger in range and scale, and much cheaper to get hold of.

Most simply perhaps, secondary data might be used to help select the sample for a further in-depth study. For example, Gorard and Rees (2002) used the electoral register in each region to select cases for their door-to-door household study of patterns of lifelong learning. The ten-yearly census of population can be a useful way of characterising the population of different regions, and so selecting an areal sample (Gorard and Selwyn 2005, Selwyn et al. 2006). The annual schools census in England can be used to select schools to represent the range of pupil intakes (Gorard et al. 2003). In all of these ways, the large-scale dataset can be argued to yield cases for more detailed study that represent the larger picture. This approach overcomes some of the deficiencies of both kinds of data, providing an audit trail of generalisability for in-depth case studies perhaps (Gorard with Taylor 2004).

Secondary data can also provide the evidence for a stand-alone initial analysis. For example, if you wish to find out whether the number of applicants to study undergraduate science at universities in the UK has been going up or down in the last 10 years it is difficult to imagine that you could collect better data on this than the Higher Education Statistics Agency (HESA) already does (or via UCAS (http://www.ucas.ac.uk/about_us/stat_services/). It is possible to conduct a new analysis of these data and produce original publishable research that could be important in terms of policy and practice (e.g. Gorard 2008b). As a snapshot consider Table 1, showing four years of new entrants to undergraduate higher education in the UK, broken down by social class.

Table 1 – Percentage of all HE students by occupational class, UK, 2002-2005

Occupation	2002	2003	2004	2005
Higher managerial	19	18	18	17
Lower managerial	25	25	25	24
Intermediate	13	12	12	12
Small employers	6	6	6	6
Lower	4	4	4	4

supervisory				
Semi-routine	10	11	11	11
Routine	5	5	5	4
Don't know	18	20	20	23

Source: UCAS
Note: Don't know includes never worked, long-term unemployed (and unknown or invalid response)

This pattern changes very little over the time period shown, despite a concurrent policy emphasis on widening participation for less elevated occupational groups. There is a small decline in the proportion of students from the most elevated occupational backgrounds, but this is matched by an increase in students whose background we do not know. Does this means that widening participation had failed during that period? Perhaps students from more prosperous families became less likely to complete sections of the application form that might be deemed relevant to income, and thought, mistakenly, to affect their student grants and loans. On the other hand, the proportion of each class in the population from which these students come also changes over time. How much of any difference is due to that? To answer the first question, we might want to conduct interviews or observations of form-filling to help decide what is going on. To answer the second question, we might want to combine the simple analysis here with consideration of another dataset, such as the census of population. Neither question can be answered with the kind of data in Table 1 by itself. Table 1 raises the issues for investigation. Secondary analysis, like most research, leads naturally to further questions and study. Use of a large scale dataset is frequently the start of further investigation, not an end in itself.

Secondary data might also be used to try and ensure that you understand the problem or pattern you are investigating with in-depth data. In the late 1990s in the UK there was considerable concern about the apparent underachievement of boys. A lot of research focussed on why boys were failing, and why girls were, for the first time, ahead of boys in terms of examination and assessment. But this research appears to have been looking at the wrong research questions in a number of ways (Gorard et al. 2001). The difference between boys and girls, where it appears, is not in terms of failing. It is, or at least was, at the highest levels of attainment, such grade A at A-level (Table 2). In subjects like maths and sciences, there is actually very little difference between the results of boys and girls. There used to be a gap in favour of boys at the highest grades, despite a higher proportion of boys taking these subjects. This gap has now disappeared. So, the follow up questions include why the gaps appear only at high levels of achievement, and why they differ over time and between subjects? Is it changes in subject entry, the nature of the teaching, or the form of assessment? Once again, we need more evidence, more data. The initial secondary analysis helps us onto the most appropriate track but it represents a starting point only.

Table 2 - Achievement gap in favour of girls at each grade, A level Mathematics, Wales, 1992-1997

	A	B	C	D	E	F
1992	-7	-4	-2	-1	1	1
1993	-15	-9	-5	-3	-1	-1
1994	-5	-2	0	0	0	0

1995	-3	-2	-2	0	0	0
1996	-5	2	-2	0	0	0
1997	0	1	0	0	0	0

Note: The achievement gap for any grade is calculated as the number of girls attaining that grade, minus the number of boys attaining that grade, divided by the total attaining that grade. This is a very simple analytical procedure.

A recent study involving the PLASC/NPD combination of datasets (see above) related to a purported decline in the number of student selecting courses in the traditional sciences (Gorard and See 2008, 2009). Where does this decline start, and who is it that drops out of science? It is certainly evident after GCSE in England, and may be related to prior attainment at GCSE/Key Stage 4 (Table 3). There are clear differences in overall attainment in sciences at Key Stage 4 between students of differing backgrounds. However, these differences are no larger than and often much smaller than the differences for all subjects. Whatever the problem is, leading to the differential attainment of social, ethnic and economic groups, it is certainly not one that is specific to science. The general patterns are the same as for science. The large gap between students identified by their schools as "Gifted and Talented" and the others is expected if the identification of G&T has been even moderately successful. It is what G&T means, after all. So what is interesting here is the relatively small gap in science. Therefore, perhaps the most worrying gap in all these subjects is between students eligible and not eligible for FSM (a measure of poverty). This useful preliminary analysis involved matching and counting millions of records. But it is simple. Using software to count a million cases is no more complex than counting 100.

Table 3 – Mean capped points scores (all subjects and sciences) and percentage attaining grade C or above (maths and English), all students, KS4, England, 2005/06

	All subjects	Science subjects
Male	338	33
Female	378	34
"Gifted and Talented"	501	46
Not "Gifted and Talented"	308	33
Non-FSM	373	35
FSM	266	25
Overall	359	34

Source: NPD/PLASC

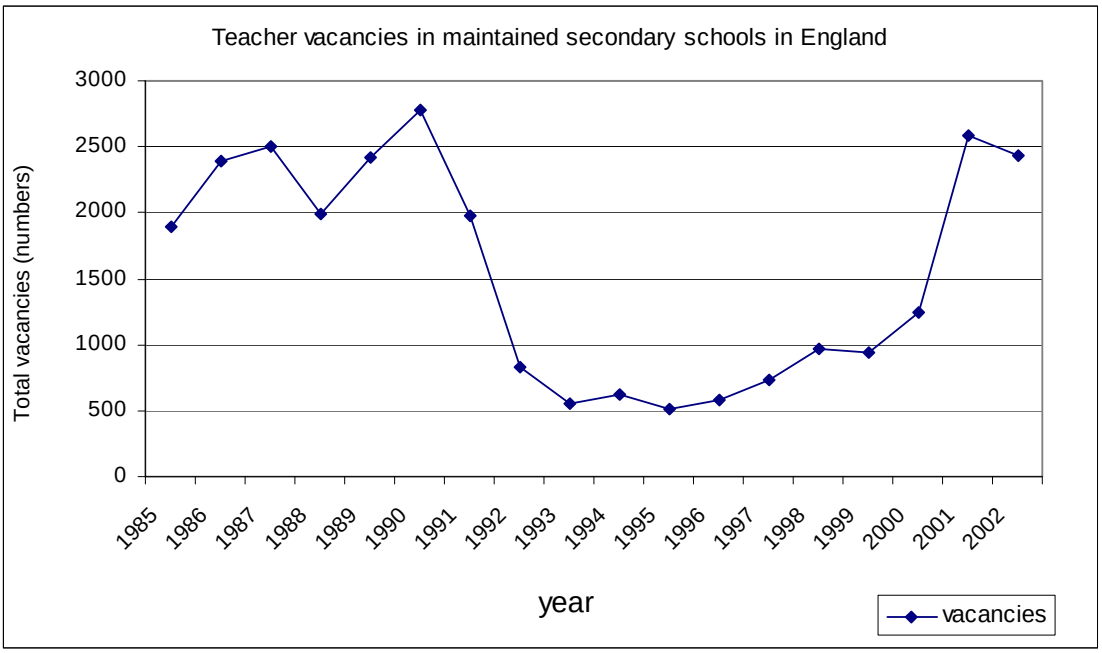
How are large-datasets analysed?

I focus in this section on the analysis of numeric datasets of the kind illustrated so far. The 'good' news for would-be analysts is that most of the secondary datasets discussed in this chapter are not based on random samples; in fact most are not samples at all but data from entire populations such as all of the pupils in maintained schools in England in 2009. This means that none of the statistical techniques based on sampling theory can or should be used with these data. No significance tests are needed or possible. No confidence intervals, and no standard errors. If female pupils attain a higher average GCSE points score than males in England in 2009 that is the

end of the matter. We can calculate and present that difference, and we can comment on it. We cannot and should not ask if that difference is statistically significant, or whether there really is a difference (Gorard 2010b). Our commentary might consider how large the difference is compared to other years, other phases, other countries, or how large it is in relation to what we know about missing data and errors in the measurement process. Again, none of these issues relates to random sampling and so no statistics, as traditionally conceived, is involved.

Here is a simple example, involving only the production of a graph, created from figures provided by a DCSF Statistical First Release (Figure 2). At the beginning of the twenty-first century, commentators in England were alarmed by an apparent shortage of teachers, especially for secondary education and in some subject areas. Newspapers, politicians and some academics talked of a 'crisis'. This 'crisis' is, presumably, represented by the rise from 1999 to 2001, when advertised jobs for teachers (vacancies) were at their highest since 1990. What caused this growth in vacancies or, looked at another way, why were vacancies so low between 1992 and 1999? Figure 2 is useful because it leads us to search for an explanation(s) that covers the two abrupt changes over time and the relatively flat picture otherwise (Gorard et al. 2006). Again, having some data leads to a desire for more. Was it a sudden surge or drop in demand caused by the birth rate of pupils? Is it a consequence of the economy, with teaching more attractive during a downturn? Was it just the consequence of increased funding – with schools advertising posts because they have the money? Was it because teachers were moving more or less between sectors such as primary, secondary, further, and extended education?

Figure 2 - Teacher vacancies in maintained secondary schools, England, 1985 to 2002



Source: See et al. (2004)

It is also not necessary to consider generic issues of epistemology or ontology when presenting an analysis of numbers like these. If you are writing a chapter based on your own interview data trying to explain why female pupils have a higher average GCSE points score than males, and you start the chapter with a table of analysis in

which you summarise GCSE points scores by sex and by subject, then the two parts are synthetic. You do not enter some different paradigm of research when presenting the table of frequencies to when you present the results of your interviews. The interviews explain and illustrate (or not) the patterns in the larger data. The common analytical themes are clarity of presentation and judgement in selecting what to present and what to omit (Gorard 2006b).

Of course, large datasets of population figures can be also modelled using regression techniques and similar, and the results can be fascinating. But such modelling is not essential and does not represent any kind of definitive test. Mostly, large datasets can be analysed as simple frequencies and/or percentages, broken down into categories such as year, sex of student, or geographic region (see Gorard 2006c for suitable approaches). It is slightly more complicated when cross-analysing two or more large datasets, but even here the complication relates to the organisation of the datasets rather than the analysis as such.

Here is a simple example of cross-pollination of two sets of figures. Academies are a relatively new kind of maintained independent school in England, the first three of which opened in 2002 (Gorard 2005). They were, at least initially, intended to improve local education in areas of high disadvantage. How have they done so far? Table 4 shows that all of these first three Academies have an increasingly different intake in terms of pupils considered to be living in poverty (and so eligible for free school meals). This change in intake should lead to higher levels of public examination attainment, since there is a well-known correlation at an aggregate level between poverty and low attainment.

Table 4 – Percentage of pupils in school eligible for FSM, 2002 cohort of Academies, 1997-2007

	1998	*1999*	*2000*	*2001*	*2002*	2003	2004	2005	2006	2007
Business	*49*	*52*	*50*	*49*	*46*	42	37	38	39	39
Greig	*56*	*42*	*43*	*31*	*39*	43	47	44	38	39
Unity	*62*	*51*	*46*	*57*	*47*	49	50	49	44	45

Note: the figures in italics are from the predecessor schools before the Academies.
Source: ASC

This is what we find (Table 5). The Bexley Business Academy had the smallest decline in FSM pupils, and shows the smallest gain in the percentage of pupils attaining level 2 (GCSE or equivalent) qualifications at the age of 16. The other two 2002 Academies had considerable increases in level 2 results - Unity from 17% in 2001 to 45% in 2007 and Greig City from 30% to 65%. This is associated negatively with a shift in FSM for both schools. Even so, these more recent gains look impressive. However, some commentators have suggested that these schools have merely changed their examination entry policies, targeting easier exams and courses. Perhaps we should test this by looking at level 2 qualifications including English and maths, the new DCSF standard threshold. And so we need to refer to more data, and the cycle of research continues. The key point to note for this chapter is that the analysis lies in cross comparison of relatively simple tables of percentages, and yet can lead to findings of national interest (Gorard 2009).

Table 5 - Level 2 percentages for 2002 cohort of Academies 1997-2007

	1998	*1999*	*2000*	*2001*	*2002*	2003	2004	2005	2006	2007
Business	*24*	*14*	*10*	*17*	-	21	34	29	32	31
Greig	*11*	*15*	*25*	*30*	-	35	26	54	59	65
Unity	*2*	*13*	*4*	*17*	-	16	17	16	34	45

Note: the figures in italics are from before the Academies. Results are not publicly available for the first year of each Academy.
Source: DCSF website

Conclusion

Whatever you wish to research relevant to education, life long and society wide, it is very likely that large datasets already exist that are relevant to your topic. These datasets are likely to be larger in scope and scale, and higher quality in terms of completeness and validity, than anything you could generate through primary fieldwork. They can be accessed directly, combined, cleaned, sorted and analysed by you making them much preferable to the third-hand accounts of evidence usually found in literature reviews.

You might use a large dataset on its own, to present a new analysis of an educational phenomenon in terms of place, time, the standard social sciences categories (such as class, sex, ethnicity, or age), or indeed any classification available from the data. You might use a large dataset at the outset of a more in-depth study – to select cases or areas, or to establish the pattern, trend or problem to be researched. More creatively, you might use one or more datasets in a synthesis with your own in-depth evidence.

It is arguably more important for you to examine the actual evidence available from previous research on your topic than it is to consider the accounts by others of that evidence in the literature. Of course, existing data, however generated and for whatever purpose, will have deficiencies. Cases will be missing, data will be missing from existing cases, measurements taken will be imprecise, some items will be miscoded, and transcription and representation may introduce further errors. The ‘construction’ of the entire enterprise may be biased, perhaps due to the underlying purpose for which the data was originally collected. But this is true of all datasets, including your own. Looking at the evidence itself gives you a good idea of these deficiencies and therefore of the substantive importance of any patterns. This is preferable to reading about the evidence in the sanitised form provided by the literature. And, of course, the limitations of any dataset can themselves be a fascinating and controversial topic for secondary research (e.g. Gorard 2008c, 2010a).

Analysing large datasets is easy, mostly because the intimidating and flawed panoply of traditional statistics is irrelevant. You do not even have to use specialist software like SPSS – Excel will do. Many large datasets are available immediately, and free for use by researchers. What possible reason could you have, except fear of the unknown, for not pursuing this? Research is, at least partly about discovery of the new. Fear of the unknown is therefore not something that a researcher can allow themselves to be inhibited by.

References

Gorard, S. (2003) *Quantitative methods in social science: the role of numbers made easy*, London: Continuum
Gorard, S. (2005) Academies as the 'future of schooling': is this an evidence-based policy?, *Journal of Education Policy*, 20, 3, 369-377
Gorard, S. (2006a) Value-added is of little value, *Journal of Educational Policy*, 21, 2, 233-241
Gorard, S. (2006b) Towards a judgement-based statistical analysis, *British Journal of Sociology of Education*, 27, 1, 67-80
Gorard, S. (2006c) *Using everyday numbers effectively in research:* Not *a book about statistics*, London: Continuum
Gorard, S. (2008a) The value-added of primary schools: what is it really measuring?, *Educational Review*, 60, 2, 179-185
Gorard, S. (2008b) Who is missing from higher education?, *Cambridge Journal of Education*, 38, 3, 421-437
Gorard, S. (2008c) Research impact is not always a good thing: a re-consideration of rates of 'social mobility' in Britain, *British Journal of Sociology of Education*, 29, 3, 317-324
Gorard, S. (2009) What are Academies the answer to?, *Journal of Education Policy*, 24, 1, 1-13
Gorard, S. (2010a) A case against school effectiveness, *Research Papers in Education*, (submitted)
Gorard, S. (2010b) All evidence is equal: the flaw in statistical reasoning, *Oxford Review of Education*, (forthcoming)
Gorard, S. and Rees, G. (2002) *Creating a learning society?*, Bristol: Policy Press
Gorard, S. and See, BH. (2008) Is science a middle-class phenomenon? The SES determinants of 16-19 participation, *Research in Post-Compulsory Education*, 13, 2, 217-226
Gorard, S. and See, BH. (2009) The impact of SES on participation and attainment in science, *Studies in Science Education*, (forthcoming March)
Gorard, S. and Selwyn, N. (2005) What makes a lifelong learner?, *Teachers College Record*, 107, 6, 1193-1216
Gorard, S. and Smith, E. (2004) An international comparison of equity in education systems?, *Comparative Education*, 40, 1, 16-28
Gorard, S. with Taylor, C. (2004) *Combining methods in educational and social research*, London: Open University Press
Gorard, S., Rees, G. and Salisbury, J. (2001) The differential attainment of boys and girls at school: investigating the patterns and their determinants, *British Educational Research Journal*, 27, 2, 125-139
Gorard, S., Roberts, K. and Taylor, C. (2004) What kind of creature is a design experiment?, *British Educational Research Journal*, 30, 4, 575-590
Gorard, S., See, BH., Smith, E. and White, P. (2006) *Teacher supply: the key issues*, London: Continuum,
Gorard, S., Taylor, C. and Fitz, J. (2003) *Schools, Markets and Choice Policies*, London: RoutledgeFalmer
Gorard, S., with Adnett, N., May, H., Slack, K., Smith, E. and Thomas, L. (2007) *Overcoming barriers to HE*, Stoke-on-Trent: Trentham Books,
See, BH., Gorard, S. and White, P. (2004) Teacher demand: crisis, what crisis?, *Cambridge Journal of Education*, 34, 1, 103-123

Selwyn, N., Gorard, S. and Furlong, J. (2006) *Adult learning in the digital age*, London: RoutledgeFalmer

Smith, E. (2008) *Using secondary data in educational and social research*, Maidenhead: Open University Press