Mixed methods research in education: some challenges and possibilities

Stephen Gorard
University of Birmingham
B15 2TT, UK
s.gorard@bham.ac.uk

Introduction

It was with great pleasure that I agreed to address the 2012 conference on mixed methods as part of UTDANNING2020 Research Council of Norway. My thesis was that what is usually called ‘mixed methods’ research in education is really just research in education. It is relatively easy to conduct, with many possibilities and few real-life challenges or barriers. What this paper tries to do is convey part of why this is so. There are of course many different methods of investigation that could be said to be ‘mixed’ in any one study – interviews with documentary analysis, or multiple regression with inferential statistics, for example (Symonds and Gorard, 2010). However, for the purpose of this brief paper, the mixture is assumed to refer to those methods that have traditionally labelled ‘qualitative’ and ‘quantitative’. For some reason, social scientists have long separated any data that involves counting or measuring from all data that involves anything else – text, conversations, observations, smells, drawings, acting, music and so on. I have no idea why. But such social scientists say that these two groups – numbers and everything else – are incommensurable, and require a completely different logic to use, and have unmatched criteria for judging research quality, and many other purported differences. Then, just to confuse things, some social scientists say that we can and should mix these forms of data – and that presumably they are not commensurable in combination, only in isolation if that makes any sense at all. It is no wonder that new researchers are confused, and that the potential users of social science evidence just ignore us. We live in a kind of la-la land.

In this paper, what I want to suggest to new researchers, and to remind more experienced ones about, is that none of the above is true. Methods are not incommensurable, and while they may legitimately be classified in a number of ways, these classifications should not become schisms. Starting with a consideration of a piece of real-life research, the paper argues that we should not separate numbers from every other form or data in the first place. Then, in terms of qualitative and quantitative data at least, we have nothing to mix. Because I do not separate the qualitative and quantitative approaches, what is termed mixed methods work just seems natural to me. It is, I contend, what anyone would do who wanted to answer any real set of research questions.

A real-life example

It is instructive to contrast how we, as researchers, sometimes behave when conducting research professionally with the ways we behave when trying to answer important questions in our personal lives. When we make real-life decisions about
where to live, where to work, the care and safety of our children and so on, most of us behave very differently from the way we do as ‘researchers’. If, for example, we were intending to purchase a house by paying most of our savings and taking out a mortgage for 25 years that is equal in size to many times our salary, then we would rightly be cautious. We would have many crucial questions to answer from the beginning, and would only go ahead with the transaction once assured that we had sufficiently good answers from what is, in effect, a serious piece of research. It is worth considering this example in some detail because it illustrates some fundamental issues about research in a very accessible way.

When purchasing a house, we will believe that the house is real even though external to us. And we will believe that it remains the same even when we approach it from different ends of the street, else why would we buy it? In these and other ways, we would un-problematically and without any trepidation just ignore the usual nonsense that is taught to new researchers as an essential preliminary to conducting research. In buying a house we would not start with epistemology, and we would not cite an ‘isms’ or Grand Theory. Nor would we need to consider the ‘paradigm’ in which we were working. We would not refuse to visit the house, or talk to the neighbours about it, because we were ‘quantitative’ researchers and did not believe that observation or narratives were valid or reliable enough for our purposes. We would not refuse to consider the size of the monthly mortgage repayments, or the number of rooms, because we were ‘qualitative’ researchers and did not believe that numbers could do justice to the social world. In other words, in matters that are important to us personally, there is a tendency to behave logically, eclectically, critically, and sceptically. We would collect all and any evidence available to us as time and resources allow, and then synthesize it quite naturally and without considering mixing methods as such. We are quite capable of judging whether the qualities of a house are worth the expenditure, for example.

If we really care about the research, as we would with buying a house, we naturally adopt what might be called a mixed methods approach. Why is it so different in academic social science then? One simple answer is that people do not care about their academic research in the same way. Another linked part of the answer is that many people purport to be doing research but in fact are doing something else entirely. I am not sure what game they are playing instead, as no one has told me the rules. But from the outside their research is similar to someone buying a house without having any idea of the price or size, or else buying it without any idea of its condition or location. Yet, education is an important applied field and the results of research, if taken seriously, can affect the lives of real people and lead to genuine expenditure and opportunity costs. So, it is quite clear that to behave like this in education research by eschewing one or more forms of data is unethical (Gorard 2002). The ‘game’ survives, I guess, simply because it is played by the majority, and so this majority also provides a high proportion of the peer-reviewers. Yet these reviewers are intended to prevent rubbish being published, public money being wasted, and education suffering in practice either by not having access to good evidence, or worse by having access to shoddy or misleading evidence.

Barriers to mixed methods
For me, that is the end of the matter really. But I know from experience that readers will want more at this stage. So, the paper continues by briefly considering some of the self-imposed ‘barriers’ to using mixed methods, and why they should be ignored. One supposed barrier, the different way in which numeric data is usually analysed, is then used as an extended example of why these barriers are self-imposed and unhelpful. The final section of the paper suggests some models or approaches to synthesising numeric and non-numeric data. There is insufficient space here to deal with every supposed barrier and every forward-looking model. What are presented instead are selected examples, with references to further published examples.

First of all, the Q words are not paradigms. Types of data and methods of data collection and analysis do not have paradigmatic characteristics, and so there is no problem in using numbers, text, visual and sensory data synthetically in combination (Gorard, 2010a). Working with numbers does not, in any way, mean holding a view of human nature and knowledge that is different from when you work with text or shapes. In the sociology of science, the notion of a ‘paradigm’ is a description of the sets of socially accepted assumptions that tend to appear in ‘normal science’ (Kuhn, 1970). A paradigm is a set of accepted rules within any field for solving one or more puzzles – where a puzzle is defined as a scientific question to which it is possible to find a solution in the near future. An example would be Newton setting out to explain Kepler’s discoveries about the motions of the planets. Newton knew the parameters of the puzzle and so was working within a paradigm. A more recent example might be the Human Genome Project, solving a closely defined problem with a widely accepted set of pre-existing techniques. The ‘normal science’ of puzzles in Kuhnian terms is held together, rightly or wrongly, by the norms of reviewing and acceptance that work within that taken-for-granted theoretical framework. A paradigm shift occurs when that framework changes, perhaps through the accumulation of evidence, perhaps due to a genuinely new idea, but partly through a change in general acceptance. Often a new paradigm emerges because a procedure or set of rules has been created for converting another more general query into a puzzle. None of this describes a schism between those working with numeric data and those working with everything else. The notion of paradigm as a whole approach to research including philosophy, values and method is a red herring. It could be argued that commentators use the term ‘paradigm’ to defend themselves against the need to change, or against contradictory evidence of a different nature to their own. They damage social science by treating serious subjects like epistemology as though they were fashion items to be tried on and rejected on a whim.

The Q words do not define the scale of a study. It has been argued incorrectly, by Creswell and Plano Clark (2007) among others, that qualitative data collection necessarily involves small numbers of cases, whereas quantitative relies on very large samples in order to increase power and reduce the standard error. But this is not an accurate description of what happens in practice. The accounts of hundreds of interviewees can be properly analysed as text, and the account of one case study can properly involve numbers. Also, issues such as sampling error and power relate to only a tiny minority of quantitative studies where a true and complete random sample is used or where a population is randomly allocated to treatment groups. In the much more common situations of working with incomplete samples, with measurement error or dropout, or involving convenience, snowball and other non-random samples and the increasing amount of population data available to us, the constraints of
sampling theory are simply not relevant (see below). The supposed link between scale and analysis is just an illusion.

The Q words are not related to research designs. What all rigorous research designs, and variants of them, have in common is that they do not specify the kind of data to be used or collected (Gorard 2013). No kinds of data, and no particular philosophical predicates, are entailed by common existing design structures such as longitudinal, case study, randomised controlled trial, or action research. A good intervention study, for example, could and should use a variety of data collection techniques to understand whether something works, how to improve it, or why it does not work. Case studies involve immersion in one real-life scenario, collecting data of any kind ranging from existing records to ad hoc observations. The infamous ‘Q’-words of qualitative and quantitative, and mixed methods approaches are therefore not kinds of research design. A study that followed infants from birth to adolescence, weighing them on 1st January every year, would be longitudinal in design. A study that followed infants from birth to adolescence, interviewing their parents about their happiness every year, would also be longitudinal. A study that did both of these would still be longitudinal, even though some commentators would distractingly and pointlessly categorise the first study as ‘quantitative’, the second as ‘qualitative’, and the third as ‘mixed methods’. In each example the design – ‘longitudinal’, or collecting data from the same cases repeatedly over a period of time – is the same. This illustrates that the design of a study does not entail a specific form of data to be collected, nor does it entail any specific method of analysis; nor does any method require a specific research design. These points are quite commonly confused in the literature, and even in many research methods resources. Such writings contribute to widespread misunderstanding of study design issues and their relationship to subsequent choice of methods. I wonder whether this confusion is sown deliberately to help the games-players evade the need for design in their own research, or to excuse their use of only qualitative methods.

One approach is not intrinsically more objective than another. Qualitative research, so its proponents argue, is supposed to be subjective and thus closer to a social world (Gergen and Gergen, 2000). Quantitative research, on the other hand, is supposed to help us become objective (Bradley and Schaefer, 1998). This distinction between quantitative and qualitative analysis is exaggerated, largely because of widespread error by those who do handle numbers (see below) and ignorance of the subjective and nature of numeric analysis by those who do not (Gorard, 2006). What few seem to recognize is that the similarities in the underlying procedures used are remarkable. Analytical techniques are not generally restricted by data gathering methods, input data, or by sample size. Most methods of analysis use some form of number, even if only descriptors such as ‘tend’, ‘most’, ‘some’, ‘all’, ‘none’, ‘few’, rare’, ‘typical’, ‘great’ and so on. A claim of a pattern or relationship is a numeric claim, and can only be so substantiated, whether expressed verbally or in figures (Meehl, 1998). Similarly, quantification does not consist of simply assigning numbers to things (Gorard 2010b). Personal judgements lie at the heart of all research – in our choice of research questions, samples, questions to participants and methods of analysis – regardless of the kinds of data to be collected. The idea that quantitative work is objective and qualitative is subjective is based on a misunderstanding of how research is actually conducted.
The underlying logic of analysis is not different. The methods of analysis for text, numbers and sensory data are largely the same, consisting of searching for patterns and differences, establishing their superficial validity and then trying to explain them. Other commentators and methods resources may claim that there is a fundamental difference between looking for a pattern or difference in some measurements and in some text or observations. This unnecessarily complex view is based on a number of widely held logical fallacies that get passed on to new researchers under the guise of research methods training. I examine one of these very widespread errors in more detail.

A logical flaw in traditional statistics

At the conference, I asked the question: ‘What is the probability of being Norwegian if in this room?’ Imagine that I was the only non-Norwegian among 100 people at the conference. Then the conditional probability of being Norwegian if in the room (pN|R) would be 99%. Anyone picked at random from the room would turn out to be Norwegian 99 times out of 100. I also asked the question: ‘What is the probability of being in this room if Norwegian?’ Imagine that there were 99 Norwegians in the room from a total population of five million. Then the conditional probability pR|N would be 0.00002. I asked if these two probabilities were the same, and all agreed they were not. I asked whether if we were given one percentage in isolation we could work out the other percentage. All agreed that we could not. We would need also to know the number of Norwegians and the number of people in the room in total. That is, we would need complete information.

To make sure we agreed, I conducted the same demonstration with a real bag of marbles. The bag contains 100 balls of identical size, of which 30 are red and 70 are blue. If someone picks one ball at random from the bag, what is the probability it will be red? This is a good example of a mathematical question that might appear in a test paper, and that has some applications in real-life, in gaming for example. We have perfect information about the size of the population of balls (there are 100), and the distribution of the characteristics of interest (30:70). Given these clear initial conditions it is easy to see that the chance of drawing a red ball from the bag is 30/100 (30%). It is almost as easy to see that the chance of drawing two red balls one after another (putting each back after picking it) is 30/100 times 30/100 (9%), or that of drawing two red balls at the same time is 30/100 times 29/99 (nearer 8.8%). Most people at the conference could either do these calculations or could see how they were possible.

Now consider a rather different problem of probability. The bag contains 100 balls of identical size, of two different colours (red and blue). We do not actually know how many of each colour there are. If someone picks a red ball at random from the bag, what does this tell us about the distribution of colours in the bag (beyond the fact that it must have originally contained at least one red ball)? It seems to tell us very little. There could be 30/100 red balls, or 70/100 or 99/100. The drawing of one red ball does not really help us to decide between these feasible alternatives. We certainly cannot use the existence of the red ball to calculate probable distributions in the population, because we do not have perfect information (unlike the first example). Yet this situation is much more life-like in being a scientific problem rather than a
mathematical one. In social science we rarely have perfect information about a population, and if we did have it we would generally not bother sampling (because we already know how many balls are of each colour). The more common situation is where we have information about a sample (the colour of one or more balls), and wish to use it to estimate something about the population (all of the balls in the bag). No one in the audience was able to tell me anything secure or interesting about the balls remaining in the bag, under these conditions.

Put into the same terms as the first example, the conditional probability of drawing a red ball from the bag if there are 30 balls in the bag (pR|30) is nothing like the probability of there being 30 red balls in the bag if we pick one (p30|R). As in the first example, one could be large (99%) and the other very small (0.00002), or vice versa, or something in between. In the usual condition of research, rather than mathematical puzzles, where do not know the number of red balls in the bag, the first probability is of no help in calculating the second. The audience agreed.

Yet, there seems to be almost a world-wide conspiracy to pretend that none of this is true when we conduct statistical analysis (Gorard 2010c). When social scientists conduct a significance test, they assume an initial condition about the prevalence of the characteristics of interest in the population and then calculate, in much the same way as for coloured balls, the probability of the observing the data they do observe. The calculation is relatively simple and can easily be handled by a computer. The analyst then knows, if their assumption is true, how probable their observed data is. For example, if they assume that there is no difference (the nil null hypothesis) between the scores of two groups in their population of interest, it is relatively easy to calculate the probability of achieving any level of apparent difference in a random sample of any size drawn from that population. This is the probability of the data given the null hypothesis (pD|H), and is what significance tests like t-tests compute. But who would want to know this figure? What the analysts really want is pH|D, the probability of the null hypothesis being true given the data they collected. As above, this is a completely different probability to the first. One could be small and the other large, or vice versa.

Yet statistical analysis as reported in education routinely confuses the two, by assuming that pD|H provides a good estimate of pH|D. So, the ‘logic’ goes, if pD|H is quite small, then pH|D must be also. But it is not true that a small value for pD|H must mean a small probability for pH|D. This step in significance testing is an error, and it remains an error however low pD|H is. The whole practice of significance testing from that stage on is incorrect and invalid. And this is true of all tests, and all other sampling theory derivatives, including standard errors, confidence intervals and complex modelling based on significance scores. Sampling theory itself, and the calculations derived from it, are not the problems here, as long as we are interested in pD|H. But no one is interested in that. As soon as we pretend that pD|H is equal to or closely related to the much more interesting pD|H, we have left the world of social science for that la-la land again.

Unfortunately for researchers there is no simple, push-button, technical way of deciding whether a difference or pattern observed in a sample would also hold for the wider population. But it does not really matter. We do not select random samples, or randomise cases to groups, in order to use statistical tests later. That would be like
saying we use crockery when eating so that we can do the washing up later! We randomise in order to try and obtain an unbiased distribution of unknown variables, as well as measured ones, in the sample. If we have randomised in order to obtain unbiased sample(s), then we could later calculate \( pD|H \) (as above). But this is a largely fruitless exercise, partly for the reason already given, but also because it does not answer the key question that is common to all analyses. This is: ‘Is the difference, pattern or trend, large enough to be worth pursuing?’. This is the same question we would ask if we had population data, no sampling was involved, and we knew the population distribution without calculation of probabilities. It is also the same question we would ask if the sample(s) did not fit the requirements of sampling theory – where the sample is non-random in nature, or where there is any non-response or measurement error, for example.

It is clear that, for any dataset, dividing the cases into two (or more) sub-groups will rarely yield exactly the same scores on all measures for both groups. It is unlikely \textit{a priori} that the school pupils sitting on the left hand side of a classroom will have exactly the same average height as those sitting on the right. Their parents are unlikely to report drinking exactly the same average number of cups of coffee every day, and so on. A difference in scores or observations may, therefore, have no useful meaning at all. Whether a difference is more than this, and is actually substantial and worthy of note, can depend on a number of factors. It depends on the size of the difference in relation to the scale in which the difference occurs (an observed difference of two feet may be important in comparing the heights of two people, but not in comparing flight distances between Europe and Australia). It depends on the variability of all of the scores. It is harder to establish a clear difference between two sets of scores that have high levels of intrinsic variation than between scores in which each member of each group produces the same score as all other members of that group. The noteworthiness of a difference may also depend upon the benefits and dangers of missing a difference if it exists, or of assuming a difference if it does not exist.

All of these issues of scale, variability and cost are relevant even if the scores are measured precisely. But in reality, scores are seldom measured precisely, and common measures like test scores, self-esteem, aspiration, occupational class and ethnicity will be subject to a very high level of measurement error. Measurement error is nearly always a bias in the scores (i.e. it is not random). People who do not respond to questions accurately (or at all) cannot be assumed to be similar to those who do. Children for whom a school has no prior attainment data cannot be assumed to be the same as everyone else. A ruler that is too short and so over-estimates heights will tend to do so again and again, uncompensated by any kind of random under-estimates to match it. Even human (operator) error has been shown to be non-random, in such apparently neutral tasks as entering data into a computer. So knowledge of the likely sources of error in any score, and an estimate of the range of measurement errors, is an additional and crucial part of deciding whether a difference between groups is big enough (to justify a substantive claim). The harder it is to measure something, the larger the errors in measurement will tend to be, and so the larger the difference would have to be, to be considered substantial. We cannot specify the minimum size needed for an effect, nor can we use standardised tables of the meanings of effect sizes (Gorard 2006). Those tables showing an effect size of 0.2 as ‘small’ and 0.8 as ‘big’ and so on are a guide only. But we can say with some
conviction that, in our present state of knowledge in social science, the harder it is to find the effect the harder it will be to find a use for the knowledge so generated. We need to focus our limited social science funding on developing effects that are big, sustained or have a high benefit:cost ratio.

Models for ‘mixing’

The extended discussion of the flaw in statistical testing is just one example of the kinds of supposed barriers we have created to hinder ourselves in the collection and analysis of different types of data. Shorn of error, the logic of analysis using numeric data involves judgement of scale, variability, persistence, accuracy, and so on, laid bare for others to follow. This is the same logic as is used, or should be used, for all data. Similarly, the other purported barriers to treating different data in a similar way are false, but there is insufficient space to view them all here (see Gorard with Taylor 2004). Of course, this does not mean that different kinds of data are not differentially suitable for different tasks. Consider the simple paper by Gorard and See (2011), for example. It uses a large-scale dataset to establish a pattern, and then tries to explain the pattern using in-depth data drawn from a sub-set of the same participants as in the large-scale dataset. Typically, large-scale data (perhaps already existing from official sources) is used to define a problem, pattern, trend or difference. It is also used to select a representative subset of cases for in-depth research to investigate the reasons for the problem, pattern, trend or difference. The in-depth work is, therefore, generalisable in the sense that this term is traditionally used, and different datasets are used to define the pattern and its determinants. This is just one of a range of simple ways in which data of different types can be used in co-operation. Others include design-based approaches (design experiments), Bayesian synthesis (that also allows the inclusion of factors like professional judgement), new political arithmetic, and complex interventions. Again see Gorard with Taylor (2004) for others.

More basically, I wonder what the schism advocates do when synthesising the existing evidence base at the outset of any new project. When reviewing literature, do they just ignore any work not conducted by people clearly within their own camp? It seems so. They do not critique the other work in detail or show why it does not meet some specified inclusion criteria. In fact, there are usually no published inclusion criteria. The reviews, such as they, are usually very partial (meaning both incomplete and heavily biased). Ideally a synthesis is an inclusive review of the literature both published and unpublished, coupled with a re-analysis of relevant existing datasets of all kinds (including data archives and administrative datasets), and related policy/practice documents. It is impossible to conduct a fair appraisal of the existing evidence on almost any topic in applied social science without drawing upon evidence involving text, numbers, pictures and a variety of other data forms. Anyone who claims to be conducting even the most basic literature review without combining numeric and textual data is surely misguided. For more on this, see Gorard (2013).

Conclusion

I wonder also if schism advocates are happy for potential research users like governments and practitioner bodies to adopt the same approach by recognising
evidence of only one kind or another? I suspect not. In fact, in the US when the government mandated the preference for funding randomised controlled trials, most academic research departments complained vociferously. They were right to complain, because a full programme of genuine research requires a wide variety of designs and forms of evidence. However, they were wrong to do so by claiming that ‘qualitative’ work was in a minority, under threat, and the only work they were prepared to do. This is blatant hypocrisy. In fact, it was probably this kind of schismatic thinking that encouraged the US government to use legislation rather than incentives in the first place.

It is not clear why everything involving numbers is counted as one approach, and everything else including smells, drawings, acting, music and so on is treated as an alternate monolith called ‘qualitative’. If researchers do, or should, naturally use whatever methods they need to answer their research questions, then there is no methods schism, and so no separate elements to be ‘mixed’ If a researcher really cares about finding something out that is as robust as possible, they should consider ignoring the traditional two-camp research methods resources and behave in research as they would in real life. In real life, the use of mixed methods is natural – so natural, in fact, that we do not generally divide data in the first place. The question to be asked, therefore, is why research should be any different?

At present, the quality of social science research in education is threatened by widespread errors of the kind reported in this paper. Reviews of evidence, and the engineering of findings into usable forms, are often impoverished by adherence to a meaningless tradition of dividing data into the two Q word silos. This is unethical from the perspective of the funders of research, and that of the general public who will be affected by the results of research. There are no real challenges to mixing data of all kinds, except the barriers that we have created for ourselves. But these barriers are insubstantial and will fall simply through us ignoring them. We need therefore to remind existing researchers how they would behave if they wanted to find something out in real-life and actually cared about the results. We also need to prevent new researchers from being taught errors in their increasingly compulsory methods development courses. This is the approach being pursued in my UK ESRC-funded project on design as the basis for analysis (http://www.birmingham.ac.uk/research/activity/education/projects/quantitative-methods-teaching.aspx), of which one of the first products is the book - Gorard, S. (2013) Research Design: Robust approaches for the social sciences, London:Sage.

References


