There is Only Research: The Liberating Impact of Just Doing Research

Stephen Gorard and Nadia Siddiqui

School of Education, Durham University, Durham, England

Overview

This article presents some higher level lessons we have learned from a full and wide-ranging programme of research. It summarises what we know about the process of research, and then looks at a number of cases—research projects in which the research questions led to the study design, and different forms of data were used in a naturally complementary fashion. The article concludes with some general observations about the liberating nature of such research for us and for our audiences.

The Research Process

The research process, as revealed by conducting many projects and reviewing perhaps 60,000 research reports over 5 years, seems to us to be both more sophisticated but less inhibiting than is commonly portrayed (e.g., Gorard, 2012; Gorard, See, & Morris, 2016; See & Gorard, 2015). There is clear spiral of research within programmes, fields, or topics moving from consideration of what we already know (Phase 1 in Figure 1) to developing ideas and artefacts, then to providing robust evidence, and eventually monitoring the results as they are used in policy or practice (Phase 7). These stages, in turn, often generate new ideas and questions leading back to the start of the spiral. For more on this see Gorard (2013).

Each phase may address a different set of research questions, from ‘what do we know about an issue?’ to ‘how can we best improve a given situation?’ And each question will involve different types of research—in the sense that the former is descriptive, and the latter is causal in nature, for example. Each of these types of research will involve different research designs. We might use a comparative design for descriptive work, a longitudinal approach to identify risk factors for an undesirable educational outcome, and a quasi-experimental approach to attempt to modify those risk factors. A crucial point to note is that none of the issues mentioned so far are related to specific methods of data collection or analysis. A longitudinal study is longitudinal whether it involves collecting observations, survey responses, or physical measurements, or all three at once. And it does not matter whether those survey responses, for example, are collected face-to-face, using Skype, by mail, or on-line. Some methods might turn out to be more useful or more frequent in some
phases than others, but generally data from several sources are collected and used in parallel in any phase—and this is illustrated later in our article.

Figure 1. The full research cycle/spiral. Source: Gorard (2013).

Having reviewed so many research reports, we can confirm that much that is published as research is not actually research at all (although some of it is useful). Much of the remaining published research is so poorly described that it is not possible to assess how much trust to put in the results. Of the minority of work that consists of research with at least adequate reporting, much of that has serious flaws. Commonly, little or no attention is paid to research design, and so inappropriate designs are linked to the research questions—comparative claims with no comparison group data, correlational and longitudinal results presented as causal, and so on (Gorard, 2015). Further problems include having a scale (number of cases per comparison group) that is clearly too small for trustworthy results, missing data inevitably leading to bias, and poor data quality. None of these issues is related to the methods of data collection and analysis used. A study comparing two groups with only four cases per cell, or a larger study with 70% attrition, generally cannot be trusted to provide good results. It does not matter what kind of data (rather than its quality) is involved.

The decision as to which data to collect is, as with preceding steps, reliant on the research question(s). In order to collect it, and to analyse it once collected, there are some craft tips that it is worth finding out about. But the general approach to handling data is very similar whatever its type. Our reviews have only found researchers wanting to portray a pattern (including uniqueness), trend, or difference. And for each of these cas-
es, the logic is the same whether it is based on texts, hospitals, socio-economic groups, or music that is being compared.

**Robust Evaluations**

In recent years, we have also conducted a large number of robust evaluations of education policy and practice, partly because of the growth of the Educational Endowment Foundation (EEF) in England, which is similar to the U.S. Institute of Education Science (IES). These evaluations have ranged from pilots to effectiveness trials, although the majority have been large efficacy trials (Gorard, See, & Siddiqui, 2017). The purpose of these efficacy trials is to assess whether an intervention (e.g., programme, process, artefact, theory) works in the sense of improving one or more outcomes, to check that in doing so, it has no undesirable unintended consequences, to estimate the extent of the improvement and its costs, to identify possible barriers and facilitators to implementation, and to provide formative advice on the intervention and any training that goes with it. This work is probably best classified as being in Phase 6 of the full research cycle (see Figure 1). As can already be glimpsed from its purposes, an efficacy trial has a wide range of research questions and leads to a wide range of data being collected, analysed, and synthesised. And, despite differences in emphasis, this is also true of pilot and effectiveness trials.

Perhaps the most high-profile part of any trial is its headline finding on impact—does the intervention work? An example in education could be impact assessment of an educational intervention on students’ academic outcomes. We can assess through a suitable design such as randomised control trial or regression discontinuity whether students have received any benefit from the implemented intervention, the magnitude of any impact, and whether the effect is positive, negative, or neutral in terms of the desired educational outcomes. And these outcomes could be of any kind and based on data of any type—the design itself is neutral on this. We have recently conducted several trials with vignettes as the outcome of interest, for example, Gorard, See, et al. (2016).

In addition, we also address more formative and contextual research questions such as:

- How well was the intervention implemented by the treatment group?
- What was the frequency and regularity of the intervention?
- How good were the training and resources for the intervention?
- Are there any barriers to implementation?
- Was the intervention subverted in any way?
- Were there any unintended consequences?
- Did the participants appreciate the intervention?
- Did this appreciation vary between different groups of students?
- What happened to the control group during the period of the intervention?
- Is there any evidence of diffusion such that the treatment became more widely used (outside the treatment group)?

As might be envisaged from these example questions, the range of evidence collected is very wide. Our approaches included secondary data, organisation websites, interviews with stakeholders, repeated observations of lessons and training sessions, surveys, and focus groups. The impact results have been compared to a black box, where results are wrapped in information, which needs opening up for developing systematic and deeper understanding of any achieved effects, and the possible mechanisms involved (Harachi, Abbott, Catalano, Haggerty, & Fleming, 1999; Oakley, Strange, Stephenson, Forrest, & Monteiro, 2004). These complementary designs and forms of data are, as far as we are concerned, just the natural way to do research.

**Aggregated trials.** Several of our efficacy trials have actually involved a bit more than this. They were aggregated trials—a relatively common concept in the field of medicine, epidemiology, and public health sciences (Chen et al., 2016). Schools and small groups of schools applied for funding to evaluate literacy interventions that they were keen to try. Their applications were refused because the scale was insufficient. However, several such groups wanted to do something so similar that they were funded to conduct the evaluation as part of a larger aggregated study, as long as they agreed to use the same timing, protocols, and outcomes. Thus, the schools would conduct the trials themselves, conducting randomisation of pupils to groups, checking implementation (as above), and leading the testing and ideally the reporting. Our role was to train the schools in running their own fair trials, oversee the work where needed, aggregate the randomisation and the test results, and, most importantly, to conduct in-depth work on how good schools are at conducting trials for themselves. Not much was known about aggregated trials in education in the United Kingdom, but if successful,
they could be large, cost-effective, and ongoing. So, as ever, the trials collected and involved the use of a wide range of data. We discuss two in more detail here—evaluations of Fresh Start, and Accelerated Reader.

Both aggregated trials used a simple waiting list design, with pupils at risk of low literacy identified and individually randomised to immediate intervention (the treatment) or delayed intervention (the control). The main outcome for treatment impact was GL Assessment’s New Group Reading Test (NGRT), with Hedge’s effect sizes computed based on comparing the two groups after the treatment group had completed their intervention. Before randomisation, the school research leaders attended a workshop on running their own trials, and where a pre-test was used, it was also administered before randomisation. The process was evaluated at each step. We conducted participant observation in teacher training sessions, observed complete sessions of the actual interventions, used teachers’ logs as secondary data to assess the regularity of pupils’ attendance in the sessions, and read the Fresh Start and Accelerated Reader developers’ website materials to understand the mechanism of interventions. We conducted face-to-face interviews with staff members, pupils, and all project leaders. The observations were simple, integrated, and as non-intrusive as possible. The interviews and field notes were part-transcribed and shared among the evaluation team members.

Accelerated Reader. This study was based on 349 Year 7 (age 11) pupils (see Siddiqui, Gorard, & See, 2016), one half of whom were randomised to the Accelerated Reader (AR) programme—a computerised intervention that assesses students’ initial reading level, suggests texts for reading that best matches a student’s abilities and interest for reading, assesses their progress, and increases the challenge over time. The treatment and control groups were found to be well balanced in terms of prior Key Stage 2 results in English. The effect size associated with the post-tests was +0.24.

In general, our observations and interviews suggested that the trial was conducted well by school and research leaders, and we concluded that schools can conduct their own robust evaluations with suitable but minimal guidance and support. There was no developer involvement. The school simply purchased the software, purchased tablets in some cases, cascaded the training, identified eligible (struggling) pupils, collated prior attainment and context data, randomised the pupils, conducted the intervention, administered the post-test, and summarised the results. As earlier, our in-depth work involved an examination of the intervention and its implementation, and these results are part of the main report, but for the purposes of this paper, we focus on the aggregated trials aspect.

Schools were particularly good at organising the intervention from the outset, arranging training, gaining permission from parents to innovate, accessing data, and monitoring attendance and progress. The latter three points are intrinsically easier for schools than for external researchers, and are part of the major advantages of school-led trials.

The weakest elements in school conduct of the trial were randomisation, handling missing data, and reporting results. The training represented an attempt to instill the idea of equipoise in the research leaders’ approach—the idea that it is more important to get a trustworthy answer than what that answer actually is because we do not yet know whether the intervention works. For some reason, schools were already, or quickly became, enthusiastic about the interventions, perhaps leading to more focus on the intervention than the evaluation (which was what they were funded for). Nevertheless, some staff understood the idea of a fair test, and we believe that with further work this element can improve.

During school visits and interviews with school leaders, we discovered that one school had allocated groups to treatment and control without individual randomisation. They had subverted the process, and explained that issues in the school timetable would not allow flexible implementation of AR in different groups. The rigour of the evaluation was not the school’s priority. This was even more obvious and general when we recommended that schools pursue pupils who had left for another school during the intervention or were absent on the day of the post-test. Many staff clearly did not see why this mattered, when the clear majority of pupils had been tested successfully. A few calculated effect sizes, but, in general, once the intervention had been completed, the schools’ new focus was on the waiting-list group and rolling out the idea, regardless of whether it was shown to work or not.

Fresh Start. This study involved 433 Year 7 (aged 11) pupils (details in Gorard, Siddiqui, & See, 2016), one half of whom were randomised to the Fresh Start (FS) phonics reading programme—a small-group, highly structured approach to teaching literacy through systematic phonics. The intervention lasted for 20 weeks and involved taking pupils outside the mainstream classroom for sessions. Using gain scores after the post-test, the intervention showed a positive impact on reading comprehension (+0.24). However, unlike Accelerated Reader, the two groups were clearly unbalanced at the outset, with weaker readers in the treatment group. This reduces the trustworthiness of the findings. Why did it happen?

Again, there was evidence of schools subverting the randomisation to some extent. The difference at the outset could have been due to chance (this is an important part of what chance means), but our open-ended
interviews yielded insights into the reasons. Unlike AR, this trial had training and input from the intervention developers (something we have found to bias evaluations, making them much less trustworthy). Here, the developers emphasised in training the importance of selecting pupils at high risk of failure in literacy. Some research leaders were enthusiastic for the weaker students to receive the intervention first. In at least one school, when pupils in the treatment school left during the intervention, they were replaced (not at random) by pupils from the control group. More work is needed to help all staff understand the priority of the evaluation over other issues if trustworthy knowledge about improving life chances is to be generated. It is possible that randomisation is one part of the process that cannot be left to schools.

Again, schools were good at handling parents, providing data, and running the intervention. Teachers were willing, enthusiastic, and well-supported by their school leaders, and they attended both workshops on how to conduct a trial. Some even managed to conduct complex school-level analysis for their own use and information. This is very encouraging.

All of the weaknesses were again linked to not keeping the fairness of the evaluation as a priority. Once teachers and assistants had been trained to use the FS approach, there were concerns that they would use this same knowledge and practice with pupils in the control group. The leaders assured us that this would not happen. But in our observations, we found FS phonics charts displayed in a classroom used by pupils from the control group. The teacher reported that it was just a visual used for teaching students in the treatment group during other sessions. Schools will therefore need more help in distinguishing between an evaluation, which is what they had all applied to do, and normal acceptable or even desirable practice. Teachers are not alone in this, and our reviews of evidence suggest that even professional researchers struggle to have the necessary equipoise needed for research.

Other Examples

There is limited space here to discuss other examples of complex interventions (as described earlier) or alternative studies that we have conducted. But we mention two more areas of ongoing research to help emphasise the point that research routinely involves collecting and analysing all and any data that could address the research question being asked.

In studies of the impact of increased school choice in the United Kingdom, and subsequent changes to the ways in which school places are allocated, we used national figures on the nature of school intakes, content analysis of admission authority brochures and websites, interviews with families, schools and adjudicators, maps and Geographic Information Systems (GIS), among other things (Gorard, Taylor, & Fitz, 2003). We therefore had secure patterns of the changes in school intakes over time, and could begin to explain these using the other sources of data. This is distinct from commentators using only one or two sources, and attempting to explain a pattern that does not exist, for example. They frequently end up using their in-depth data for unwarranted claims about national patterns, or else attributing a role for choice in explaining the aggregated patterns that is not justified by the data they collected.

Similarly, when attempting to explain why some adults seem to be progressively excluded from formal opportunities for education and training, we used secondary (census) data, large-scale surveys, household interviews, case studies, oral archives, and local histories. Some of the publications focused on the patterns (Gorard & Selwyn, 2005), others on the stories (Gorard, Rees, Favre, & Welland, 2001), but mostly reported them in an integrated way as reliant on each other (Gorard & Rees, 2002). Using large-scale data tends to emphasise the role of structure and even predictability at an aggregated level in people’s lifelong trajectories. The same factors rarely appear in individual’s in-depth accounts, which tend to emphasise choice, the role of others, and serendipity. As earlier, using only one of these forms of data would be likely to impoverish and even bias the findings, and any practical conclusions drawn from them.

Discussion

Readers will have noted that this article has so far avoided use of terms like mixed or multiple methods, or the q-words qualitative and quantitative. That is, because we do not like to use them except to explain why we do not use them (or where we are occasionally forced to by circumstances). The q-words are at best redundant, and at worst divisive and misleading. Some simple research questions, such as how many students of a particular ethnic group started university in a country last year, will have simple numeric answers. Other questions, such as what a group of teachers think about a new policy, will have narrative answers. This is clear, unproblematic, and does not require a paradigm or specific approach to research. There is no such thing as quantitative-
tive or qualitative or mixed methods designs. In most real-life curiosity-driven research, the research question or questions entail a wide range of data being collected and synthesized, as exemplified in the earlier studies. Such studies could be called multiple methods but this term adds nothing to our understanding of them, any more than saying a numeric answer is quantitative or a narrative qualitative. If methods are used naturally in combination as the scenario demands, then there is no need to mix methods because there has been no separation in the first place.

We apply this idea to our teaching, research capacity-building, knowledge transfer, and all of our research. It makes the work simpler, the communication of findings easier to wide audiences, and it makes reviews of evidence more secure and less biased. We are not sure how researchers who persist in believing in paradigms and so on manage to conduct literature reviews. Do they totally ignore (or meekly accept) work from what they see as another paradigm? If so, this is a far worse problem than the file-drawer issue or even the misguided use of scale and quality to eliminate some studies from any consideration when reviewing (Gorard, 2015).

We find this approach easy to convey, and less inhibiting as researchers than having lots of rules to follow about what one can and cannot do with any design or combination of data. Then, there is only research.

References


