Interpreting Statistics in an English Team-Based Evaluation[1]

PETER BOWBRICK

Contents

[1] Published in International Handbook of Interpretation in Educational Research Springer International Handbooks of Education 2015, pp 1347-1380 link at http://link.springer.com/chapter/10.1007%2F978-94-017-9282-0_66

ABSTRACT

This chapter describes an evaluation carried out by a team. The technical problems to be tackled in interpreting statistical data are discussed. It is shown how the statistical analysis is then interpreted by the team as a whole to produce a consistent and coherent report.

The objective of the evaluation was to influence government policy.

The political and micro-political pressures are examined. The way in which readers of the report drew inferences from the report, and interpreted the statistical results, and then disseminated them to influence policy is discussed. This raises serious ethical issues.

INTRODUCTION

> *'Teachers – except, of course, statistics teachers – sometimes commit the regression fallacy in comparing grades on a final examination with those on a midterm examination. They find that their competent teaching has succeeded, on the average, in improving the performance of those who had seemed at midterm to be in precarious condition. This accomplishment naturally brings the teacher keen satisfaction, which is only partially dampened by the fact that the best students at midterm have done somewhat less well on the final – an 'obvious' indication of slackening off by these students due to overconfidence.' (Wallis, W.A. and H.V.Roberts,* ***Statistics – a new approach,*** *Methuen, London 1957*.)

THE STUDY

A team consisting of two educationalists, Professor Morwenna Griffiths and Dr Tony Cotton, with me to do the statistics, evaluated a Department of Education and Science (central government) project for improving the results of selected schools. The time line and funding had been specified by the clients, who also stated the data availability and some of the analysis that must be done. This chapter, therefore, is concerned with drawing inferences in the common situation where time and money are major constraints, where the statistical analysis is only part of the team effort, and where the message communicated to the clients is important. Certainly it would have been carried out very differently as part of a well-funded PhD, and different statistical inferences might have been drawn, but the results would have come too late to affect what the clients did.

The Intensifying Support Pilot (ISP) was designed to offer a package of support and professional development to schools that had low achievement in literacy and mathematics and had made little progress in raising standards since the introduction of the National

Literacy and Numeracy Strategies, a central government 'initiative'. From the start, the ISP had been fully integrated with the rest of the government Primary National Strategy. It aimed to draw together existing good teaching and learning practice. It was designed to raise standards and improve teaching and learning in the context of the school as a professional learning community. The programme worked in partnership with the LEA (the local government education authority which runs the schools within its area) and the school. It was based on the cycle of audit, setting targets, action and review. It was designed to support schools to establish self-sustaining systems.

The different players in the system were the main government department (the Department for Education and Science in London), the section in this department responsible for the National Literacy and Numeracy Strategy, the ISP management team working from London for this section, the LEAs, the 'consultants' (who were experienced teachers working full-time on delivering the support programme to the schools within their LEA), the schools, the headteachers, the teachers and, of course, the pupils. All had an influence on our work, and all but the pupils drew inferences from our reports.

In the pilot round of the ISP, LEAs with a relatively high proportion of schools in the lowest attaining category on a national scale were identified (May 2002). They were invited to join a programme to pilot a programme under which more intensive support would be delivered by consultants to some of the schools within the LEA. We were appointed to evaluate the pilot in July 2003. In October 2003, an extension of the programme to 76 further LEAs was started. We were then asked to expand the 2004 evaluation, first to see what happened to the schools that had been in the ISP pilot programme when the support was no longer provided, and, second, to compare the current views of the schools in the second phase with the views that had been expressed by the pilot schools.

Aims of our Evaluation

Our evaluation examined the delivery of the ISP in the first phase, identifying what seemed to have worked and what problems were encountered. We were also to identify any measurable results, and, specifically for the second (2005) evaluation, the impact on attainment in National Curriculum tests at the end of Key Stages 1 and 2. (There were state-run tests at Key Stage 1, the first year after the Foundation (infants) stage and, again four years later at Key Stage 2. Schools were judged on the percentage of pupils reaching a certain level in the Key Stage tests.)

In both stages we were required to carry out statistical surveys and do statistical analysis of test results. That was largely my responsibility

The evaluation did make substantial use of the statistical analysis to determine whether it could be shown that it had in fact improved the educational attainment of schools in the programme. However, it did have much broader objectives, which required an evaluation of changes in the teaching and administrative approaches. It required, therefore, a team approach.

THE CLIENTS' OWN SURVEY

The clients carried out their own survey before our contract started. It was intended that LEA consultants would hand out the questionnaires during a meeting with the staff of a school, and that the questionnaires would be filled in during that meeting. This was done by some consultants. In other cases the consultants sent questionnaires to the schools. In some schools nothing appears to have been done with the questionnaires. Indeed one school sent the questionnaires in, untouched. In other schools there was a small response.

This means that there are several forms of response bias. One is that schools dealing with certain consultants and certain LEAs either did not respond at all, or had a low response rate. It is conceivable that these consultants and advisers were better or worse than the average, and so the results were biased. Another source of bias is where some teachers chose not to respond. These may have been teachers who were particularly critical of the pilot programme. We do not know. The postal questionnaires were submitted later than the others, and may reflect how people felt a month or two after the end of term and were rested, rather than at the end of a busy term like the others.

It is also apparent from an inspection of the returns that responses within a given school were surprisingly similar. It seems likely that when the questionnaire was presented to a group after that group had spent an hour identifying and discussing problems and achievements, they all would write down the problems and achievements discussed during that hour.

A disproportionate number of the responses came from a few LEAs where the survey was carried out as intended. In these LEAs there were responses from more schools, and there were more responses per school.

We did not think that anything could be done to remedy this by the time we saw the data. It seemed unlikely that the normal process of sampling the non-respondents to see whether they were atypical would work. There had been a considerable delay since the survey took place, and perceptions had undoubtedly changed after the holidays. There was a high staff turnover, many of the staff would have moved on – possibly those most disaffected, possibly not. Response rates were likely to be very poor. It was felt, too, that the follow-up would be perceived as more paperwork, and would be resented. We had no doubt too that any follow up to the ISP survey would damage the mail survey we were to carry out.

One problem was that the questionnaire had not been piloted. For example, a large proportion of respondents had a problem with the question 'What is your current teaching responsibility?' This meant that it was not possible to analyse the data by whether the teacher was a full-time class teacher, a head teacher, a subject coordinator etc. It turned out that in most cases they had two or more areas of responsibility, In fact, when all the combinations were allowed for, there were 119 different categories. Any analysis by responsibility would, therefore, have been seriously misleading.

The only value of this survey to our evaluation was that some of the written comments gave an indication of what issues we might ask about.

It is relevant that our clients were perfectly happy with their survey, and suggested that the biases could be ignored because there was a misunderstanding about how the questionnaires were to be distributed.

OUR MAIL SURVEYS

We carried out three mail surveys requested. In the review of Phase I there was a mail survey of the headteachers involved in the ISP and a mail survey of the consultants delivering the ISP in the LEAs (that is to say not the ISP staff operating from head office in London). In the review of the second phase, there was a mail survey of the head teachers of the schools which joined the ISP in this phase.

I had carried out a lot of surveys when I first started in research many years ago, trying to be formally correct in carrying them out so the results would be valid. I came to realize though, that, for my type of research, surveys were an expensive way of putting numbers on what I knew already. The open-ended depth interviews which start the process of identifying what the survey is to be about and what questions should be asked usually made the survey redundant – either they had told me what I needed to know, or they had told me that the real question was something I had not thought of.

Mail surveys take a lot of time to administer. A survey would normally take place over more than four months, doing it according to the standard research protocols, so instant results are not possible. This makes them extremely unattractive when I am doing my own administration, or where I am working to a time limit. In my consultancy abroad my statistically literate employers would not let consultants do surveys. In Britain I note that some consultancy firms routinely carry out mail surveys. The advantages to them are, first, that the surveys can be carried out by cheap clerical staff, second, that the analysis can be done very quickly using SPSS, third, that few clients will know or care if costs are cut by omitting some steps of the standard survey protocols, fourth, that this analysis can produce an enormous number of tables and graphs to give the clients the assurance that a lot of work has been done, and, fifth, that publishing programmes can present the tables and graphs very prettily, if not comprehensibly. This is very impressive to clients not trained in statistics, who may think that these tables and graphs alone justify the consultancy fee.

In our evaluation none of these considerations applied. The terms of reference required that we do a mail survey.

Our survey

Surveys provide a lot of the data used for drawing inferences and there is a tendency to accept that they are grossly inaccurate, and then treat the figures as though

they are 100% correct. Sampling error is usually quoted in reports, because it can be calculated quite easily, but the other errors are usually forgotten - questionnaire bias, bad questionnaire design, interviewer bias, respondent bias, recording error, calculation error and aggregation error. We tried to bear these in mind in drawing inferences.

Response rate

Our major worry was that we might not get a big enough response to draw any conclusions. We did not know, or care, what might be acceptable to the clients. I have seen a British government department express itself delighted with a mail survey that had ten per cent response, though these ten percent were obviously completely atypical in this way at least, and a quick glance at the results showed that they held extreme and incoherent views (Bowbrick, 2012). We do not share the government's view of what is an acceptable response.

It was clear from the clients' own survey that non-response could be a serious problem, invalidating the whole study. It was clear from our team's initial interviews that the headteachers were swamped with paperwork – they were dealing with up to 30 central government 'initiatives', like the IPS, and had to handle the enormous amount of administration imposed on them from above, as well as running the school. It was also clear that the consultants running the process were swamped with paperwork from headquarters.

This meant that the questionnaire had to be very short, and obviously easy to fill in. The questions had to be clear, so there could not be the slightest confusion about what they meant. The questions had to be obviously relevant – we could not expect the questionnaire to be returned if questions were perceived to be silly. We had to leave space for the respondents to put in their own comments, both to encourage them to reply and because it was an important part of the study. We had to guarantee anonymity – our guarantee would carry more weight because we were operating from a university. These considerations were important in preparing the questionnaires and in piloting them.

The survey covered all LEAs and schools in the ISP. The questionnaire was mailed to headteachers, project consultants and LEA personnel. The questions were designed to elicit views about each of the themes and elements in ISP. Respondents were asked to answer each question on a five-point scale and were then invited to comment on the reasons behind their response.

The clients provided us with lists of all three populations together with addresses and contact details. Surprisingly, these were faulty: 6% of the schools on the list had been closed, and 12% of the consultants were uncontactable, because, for instance, they had changed jobs.

The questionnaires were sent out in reply-paid envelopes. Two reminders were sent to non-respondents. In the two cases where there had been a low response rate from an

LEA, follow-up telephone calls were also used, and the opportunity was used to ask deeper questions, addressing issues that had arisen.

In the evaluation of the first phase we got the following results

		No	%
Number of questionnaires sent out to heads		132	100
Number of replies		117	89
of which	Schools closed etc	8	6
	Unusable	4	3
	Usable	105	80

		No	%
Number of questionnaires sent out to consultants etc		49	100
Number of replies		33	67
Of which	Not contactable	6	12
	Not usable	3	6
	Usable	24	49

The responses from the initial posting, the reminder and the second reminder were compared and no significant differences could be observed. It was concluded that the response from the headteachers was satisfactory and that it could be taken to be representative of the population as a whole.

We were not happy with the response from the consultants. The response was a lot lower than I would expect, and much lower than we got from the headteachers when using the same methods. We had expected a higher response if anything, because the consultants were employed specifically to deliver the ISP, while the ISP was just one part of a busy headteacher's job. In our evaluation, therefore, we had to bear in mind the possibilities that for instance, the consultants might be so overloaded with paperwork that it was too much to spend ten minutes on a survey, or that they were extremely unhappy about their work and were frightened of making any comment that might be traced back to them: it might affect future employment with ISP or the LEA.

Generation of questions

It is standard practice to do depth interviews to find out what questions should be asked, and how they should be asked. The questions were drawn up after the team members had had in-depth discussions with people in the clients' headquarters, with field consultants and with headteachers and teachers. They were then piloted with people in the target populations.

For a later survey, depth interviews in selected LEAs, we were able to pilot in LEAs which would not be covered.

Drawing Inferences

The questions had a range of objectives.

The fundamental one was to provide crude answers for the evaluation. Did the consultants and headteachers think that the ISP was working well or not? Which elements did they think were working well or not? This provided the headline information for the Department's public relations: yes, most of them thought it was working well.

The crude figures could then be broken down in the analysis. Analysing by LEA (Local government area), for example, showed big differences between perceptions, suggesting that the stated success of the ISP depended very much on the competence and charisma of the individual consultant responsible for that area.

The first inference was that there was some reason why we had a very low response from the consultants, compared to headteachers, but we could not know what it was. Our response was for the team members to do what they could to identify reasons, using depth interviews and telephone interviews at the later, interview, stage of our study. Statistical methods were inappropriate when we did not know what questions to ask or what the perceived problems were. That is to say, the temptation to draw conclusions directly from the survey results had to be resisted.

In drawing inferences it was important to remember the importance of the charisma of individual consultants. This was stressed to us repeatedly by the ISP management and came up again and again in the team's interviews. We observed a meeting held by London-based ISP staff for the consultants which reminded us of a revival meeting, whipping up a crowd's enthusiasm. This does raise the possibility that the people who reported themselves to be strongly in favour of ISP were in fact reflecting the fact that the LEA consultant had inspired them with enthusiasm. Some consultants clearly aroused this enthusiasm with a lot of headteachers, some only with a few. We did not have the time or money to see whether those LEAs or schools which had been most enthusiastic about ISP did actually produce the best results, or whether there was any relationship between the replies of individual consultants and the results in the LEAs in which they worked. The data on results was not available in time. Some questions still remain, therefore. Would having any sympathetic listener who was willing to listen to problems have had the same effect as the ISP (as some people suggested in interviews)? Would having any person with charisma have the same effect, regardless of technical competence? The design of the ISP did not let us draw any inferences on this.

An important purpose for us, as evaluators, was to find out what aspects of the ISP were disliked by consultants and by headteachers and what aspects they did not think were working well. This was addressed both by specific questions and by the comments that the respondents were invited to give. The results could then be addressed later in our depth interviews.

There turned out to be a fundamental disagreement with the clients on this, a major difference in interpretation. They were of the view that if most people thought part of the system was performing well (as they did), we could ignore the fact that perhaps a quarter thought it was performing very badly, and we should omit any mention of this quarter from the discussion in the report. We, however, thought it important to find out more about what was potentially a major problem running through the system as a whole: it was not satisfactory that a quarter to half of the headteachers were unhappy with an element of the programme.

Partly as a result of this difference in interpretation, they were reluctant to have the critical comments published in the report, though they were happy to give great prominence to any favourable ones. (I discuss below the suppression of unwelcome results in general.) The fact that a problem had not yet impacted on most people does not mean that it never will. The fact that a problem had not yet been noticed by most people does not mean that it is not already impacting on everyone. Indeed, it has been my experience when doing evaluations and investigations that the key fact that explains all the inconsistencies is often given by just one respondent.[2] In drawing inferences from such information,

[2] I struck oil this way when I was evaluating a large irrigation project designed to plant a million apple trees in Pakistan. I got stuck in Peshawar for three days, waiting for a plane to take me out. I saw the people I had to on the first day. To pass the time I spent the next two days interviewing other people, people who might have had a different perspective. I hired a rickshaw for a day, and went from office to office. I enjoy people and this was a lot more fun than sitting in a hotel bedroom. I learnt a lot about things that interested me, though not very relevant to my project. Eventually someone let slip the fact that in the Tribal Territory of the North West Frontier, it was not the Department of Agriculture who distributed apple trees, as in the rest of the country; it was the Department of Forestry. I rushed to see them and found that they had handed out a million apple trees in the past five years. How could this be? All the project documents I had been given had masses of statistics and there was no mention of these million trees. Nobody I had met in the provincial or national Departments of Agriculture had mentioned them. It turned out that the Tribal Territories are not fully incorporated into Pakistan. They will be when they pay taxes, but this will not happen until a woman can walk from one end of the village to the other in perfect safety – which is not likely to happen for years. In the meantime, the Tribal Territories are treated as semi independent. Their statistics are not included in the Pakistan statistics, but in a separate volume.

This little bit of information blew the project out of the water. It had only appeared to be profitable because there was a shortage of apples in the country, so apples got four times the price of oranges. This extra one million apple trees, plus the million trees of the project I was evaluating would create an oversupply and push the price of apples as low as that of oranges or lower, so the whole project was uneconomic. This one interview saved $50 million directly, and hundreds of millions in wasted investment by farmers. But the result was entirely unexpected. Nobody I had met before had any inkling of this, and nor did I.

sampling method, response rates and statistical reliability are irrelevant: someone has pointed out that there is a potential problem, and we should investigate. The evidence to draw conclusions will be found elsewhere.

Most of the analysis of the surveys did not impact on the evaluation. The tables were presented in a format that would facilitate the management of ISP in future, by the ISP management and by the LEA consultants trying to find out what was happening in their own areas. It was clear that the tables would have to be presented as simply as possible for an audience not trained in statistics, and that we could have no control on how they chose to interpret them.

DID THE ISP HAVE ANY EFFECT?

The most important questions our team had to ask in the evaluation were, 'Did the ISP have any effect?' and 'If it did improve outcomes, by how much?' If it had little or no effect, it was not worth continuing with.

For the quantitative aspect, obtaining the data was the first priority. As is always the case when examining a real world situation, the team started without knowing what was relevant. The first step of our study was, therefore, to build up non-statistical models on what was happening. The team produced a literature review and interviewed civil servants, the ISP management, consultants in the field, headteachers and teachers. Once we had this information, we looked for the data to build a statistical model. We soon found that no statistics existed on many of the aspects that were obviously important and that many of the statistics that did exist were wrong, or were wrong for our study. This is perfectly normal, but it means that a lot of thought is needed before constructing any statistical model. We had to tackle these problems before we could start drawing inferences.

Working Without Data

No statistics existed for most of the key factors, on inputs or outputs. This had to be taken seriously: it meant that it was not possible to create a statistical model that reflected our model of what was really happening.

We could of course build a model using the data that did exist. There is always a danger that you may leave out a factor because there are no statistics, because the statistics are unreliable, because no observations are recorded for it (which, as will be shown later is far from saying that the factor does not exist or has not been observed), because the analysis of the statistics that are available is too difficult, or because it would make the

model too difficult. There is also the opposite temptation, to include irrelevant data because they exist. It takes a certain firmness to refuse to use some of the vast amount of information produced as a by-product of administration. There may also be a temptation to put all the data that do exist into the computer without prior constraints, to run random regressions and to include everything that produces a good 'fit' into the model. These temptations are particularly strong when one is under pressure to start the analysis as soon as possible, or when the output is an academic paper rather than a recommendation for action.

In this evaluation, lack of data was largely because it is an area of rapid change or great confusion, which is exactly why there was a particularly strong need to investigate it very carefully indeed.

In a sense, leaving out key factors because the data are not available is like removing a leg or two from a chair. Putting in factors purely because the data are available is rather like adding a leg or two, pointing up or sideways rather than touching the floor. That is to say one removes any security from the chair, the other adds expense and complications without increasing security.

Our team worked closely together to limit these dangers. The other team members told me what their work showed them were the key factors, and I told them which areas could not be covered by the statistical analysis and must be addressed in their interviews.

Non-reporting of selected observations

The dangers of an incomplete model are particularly likely to arise out of non-reporting of selected observations or of a class of observations. This usually happens because the people responsible for primary data collection fail to report certain observations, but it may happen where selected observations are removed during the data processing or analysis. An example of this bias is where scientists report the yields of a new strain of wheat on those test plots where it gives an increase in yield, but not elsewhere. They may justify this to themselves on the grounds that low-yielding plots must have been affected by extraneous factors that were not observed, that, for instance, 'They look as though they were damaged by windborne herbicide' or 'Obviously this patch was damaged by white mould'. Similar suppressions are common enough in the administrative procedures in the collection of much of the raw data of educational statistics. This is an example of deliberate suppression, though by people who believe that they are being perfectly honest.

Clearly, this suppression invalidates all the perfectly accurate observations in the experiment. The final result is wrong. It means that it will throw doubt on the correct result produced by normal methods. It also throws suspicion on all other statistics collected by the same organization. It changes the standard error as well as the mean.

In this study there was non-reporting of observations. Deliberate non-reporting to bias results is discussed below. The analysis was also handicapped by non-availability of data when they were needed. It is difficult to understand, for instance, why full information on

the percentage of pupils in a class reaching Level 4 at Key Stage 2 should be available, while the average score of the pupils in the class should not be available for months: the same data are used with a very small difference in the calculation. Nevertheless, this lack of data seriously restricted our analysis, and the inferences we could draw.

In a later section I discuss some more general problems with the use of official data available.

Non-reporting of a class of observations

The best-known example of the non-reporting of a class of observations was the Thalidomide affair. In the trials of this drug, there was full and accurate reporting of a wide range of phenomena. This showed Thalidomide to be an extremely effective and useful drug with no harmful side effects on the patient. It was only after the drug was released that it was found that it had very serious effects on the foetus of a pregnant patient. The failure to test for this, and the delay in recognizing it when it happened, was due to a weakness in the theory: it had not been realized that the foetus could be damaged by a drug administered to the mother.

There is a difference between a) failing to notice that a category, such as 'deformed babies' might be relevant, b) failing to notice that a category such as 'deformed babies' exists, c) failing to report observations in that category and d) reporting that there were no occurrences in that category, rather than that no observations have been made. In educational statistics it is not usually clear which of these causes the lack of data or the blank in a statistical series. It is not always clear whether '…' or 'n.a.' means that there was only a negligible quantity in the category or that no data were recorded. If the statistical table does not include a heading for a category such as 'deformed babies', it is hard to find out whether it is because of a), b), or c) above. It may be because someone thought the category was unimportant, or because the results looked suspect.

Working With Bad Data

Most non-statisticians ignore data errors when working with statistics. The more conscientious may quote the sampling errors, but not the other errors and then proceed to use the data as though they were 100% correct. When they receive statistics they ignore the caveats that the statisticians produced to accompany them. People with some statistical knowledge may run regressions or other models and, when they get a 'reasonable' fit take it as an indication that they were right to take these caveats and sampling errors as only an extremely unlikely theoretical possibility. It is common for them to report their results to three significant figures, accurate to one tenth of one per cent, and it is not rare to see them quote results to one part in a billion.

This does not mean that one should ignore all bad data and statistics. All data are in some way bad for any particular study, and ignoring them would mean relying on wild guesses instead, so the data and statistics must be analysed rigorously, bearing in mind their

weaknesses. This means that there are necessarily a lot of theoretical models and theoretical techniques that are useless in a particular study because they need quite more accurate data than does exist, or indeed than could ever exist or because they need quite different data. If the models are used, they produce results that are wrong but have a quite spurious appearance of reliability because of the sophisticated models used.

In my first job I had to publish a monthly bulletin of statistics. I had to go into government and semi-state organizations, get hold of any information I could and prepare it for publication. There were enormous errors due to the collection procedures and the way the raw data was processed in these organizations. The government policy was to publish them because they were the best we had. In spite of this, the figures looked completely authoritative once they were put into print. However long the cautionary footnotes I put in, the statistics were used as though they were accurate. It is surprising how few statistics textbooks give more than a passing mention of this, perhaps a couple of pages suggesting that the raw data or the published statistics may be imperfect.

All statistical organizations, public or private, have to justify their existence to non-professionals. Non-professionals expect statistics to be accurate and they are likely to cut budgets if they find that the organization is producing figures that are wrong. As a result organizations may be reluctant to admit that their figures are anything but perfect. There is a tacit agreement that statistical organizations do not criticize each other publicly – it is common for different organizations to produce completely different figures for the same phenomenon without mentioning each other's figures or mentioning that there are discrepancies, and I have seen one small organization publishing two series showing very different levels and opposite directions of movement over time for the same phenomenon, without anybody in the organization realizing the contradiction.

Researchers often use techniques that give a wholly spurious appearance of accuracy, because the non-professional does not understand that complete accuracy is impossible. For example, they quote four or five significant figures for statistics derived from data that are only accurate within 30%. This is a falsehood. In fact, Morgenstern (1963) recommended that all statistics should be presented with only the number of significant figures justified by their accuracy. He suggested, tongue in cheek, that the enormous savings in printing costs would finance a major improvement in collection procedures.

We must start, therefore, with the assumptions that

1. There is probably a large random or biased error. It can never be assumed that the error is constant, whether a constant sum, a constant percentage, or in a constant direction. Nor can the opposite be assumed, that there are random errors which will cancel each other out over time.
2. For most decisions we are not concerned with averages: we are concerned with changes at the margin, at what would happen if we put a few more resources into certain schools (as in this case) for instance. The margin is the difference between two figures, the results before and after the ISP, but each of these figures is subject to random and non-random errors. Even if the

averages for the sector as a whole were very accurate, the figures for the margin, for the year on year changes in these selected schools, would be highly variable.

As the study progressed, we discovered more and more reasons to believe that there were random and biased errors which would make statistics for changes at the margin extremely unreliable. These are discussed below.

Noise

Computers have resulted in an information explosion. There is now a vast amount of information collected, often as a by-product of an administrative procedure. The data may be made available, and statistics published, but they are of unknown provenance, reliability and error. We do not know how the data were collected or processed, nor do we know the definitions used. This is just 'noise'. It means nothing in itself, but it stops us from identifying, isolating and using the meaningful information, and from identifying gaps. Reliable and accurate statistics which are not relevant to our task must also be classified as 'noise' for our purpose. The first, and most important, part of the task of interpretation is rejecting information which is noise.

Are errors constant?

A common reaction to the knowledge that the statistics are unreliable is to say 'I know that the figures are wrong and are probably not within 20% of the true figure. However, I am not interested in the average, but the change from year to year. The average may be wrong but the changes from year are right. I can still use them for trends, regressions and correlations.' In effect, this is assuming that the error is constant. Depending on the specifications of the model, the implicit assumption may be that the error is a constant sum, a constant percentage or even a constant power. Often one sees that in one part of a single model a constant sum is implicitly used, in others, a constant percentage. The assumption is that the error is all a constant bias, or, sometimes, that the random errors cancel out. It assumes that there was no change in the bias over time. These assumptions have to be identified, and challenged, before interpretations are made.

For some reason, too, it is very easy to convince oneself that the results of an obviously unreliable survey become completely reliable if the survey is repeated year after year. Because the survey is carried out every year, researchers, including the statistically literate, have no hesitation in using the figures as being both perfectly accurate and strictly comparable from year to year. Perhaps the feeling is that the means (Arithmetic? Harmonic? Geometric?) may be wrong but the changes will be meaningful, an assumption which is false, for reasons that will be discussed later.

In our evaluation this was particularly important. Since we were carrying out the evaluation before the intervention was complete, since we were measuring changes in test scores within one or two years of the intervention, and since data were not available immediately after the tests were run, we were, in effect, looking at year-to-year changes in test score. We were also concerned with changes at the margin. For reasons set out below there were strong reasons to expect random and non-random fluctuations.

Some models are sensitive to bad data

Morgenstern (1963)quotes an example from Milne (1949)to show that slight errors in specification and in input can cause enormous errors in output. He takes two equations:-

$$x - y = 1$$
$$x - 1.00001y = 0$$

which have the solution x = 100 001 and y = 100 000

The almost identical equations

$$x - y = 1$$
$$x - 0.999999y = 0$$

have the solution x = -99999 and y = -100 000

While this is an extreme example, you cannot know whether your model will produce similar results unless you test it with variations of the data within the limits of its error. It is possible that the errors might cancel out, but it is equally likely that they would multiply. The implication is that if you feed data with a small error into a complex mathematical model you can expect to produce results with a large error. The more complex the model, the more likely it is that it will produce large errors in an unexpected direction.

This problem is often tackled by saying in the report, 'This model is robust,' and scattering the word, 'robust', through the report, a technique that reduces awkward questions but does not solve the problem.

A lot of simplifying assumptions must be made to make a theoretical model workable. However, each explicit simplifying assumption necessarily introduces implicit assumptions which usually go unrecognized. The more simplifications are made, the greater the effect and the greater the danger.

It is often forgotten that the raw data have been aggregated then further processed to produce a statistic. This mathematical aggregation and processing is part of the mathematical model when the statistic is used as an input into the model. In our study this hidden analysis proved crucial.

Available Statistics

Data availability

The interpretation is strongly affected by the data available, and it is a truism that the result can be altered by deliberate or accidental suppression of certain data.

We were astonished to find the sheer quantity of data that could be amassed by a large government department in a rich country, and even more astonished to find how little of it was of any use to this study. Most of the professional statisticians, not least those I consulted when preparing for the present study, have a strong professional integrity which compels them to make available all they know of errors and mistakes in their output. They usually present a very full analysis in the technical appendices, and they are willing to discuss the basis of the figures I want to use, and whether my use is valid. At the start of the evaluation I discussed the availability of statistics with statisticians from three LEAs and from the central government Department for Education and Science. They urged caution.

If data are collected and statistics produced by people who have not thought out clearly who is going to use the resulting statistics and what decisions they are going to use it for, it is probable that the wrong information will be collected, that it will be processed in a way that makes it inaccessible to the user, and that it is delivered long after the decision has been made (Bowbrick, 1988). This is particularly common when the data are the by-product of administration. Often the definitions and methods used for collecting them, then processing them to produce published statistics may make the statistics unusable for many purposes.

In our evaluation, the only possible measure of whether there had been any impact (as defined in the terms of reference) was the tests carried out first at Key Stage 1 and then, four years later, at Key Stage 2. If the schools that adopted IPS had better results in these tests after a year using ISP, it was an indication that ISP might have had some effect.

The Pilot Sample

There are rules for carrying out an experiment to see if a pilot programme is working. It would be normal for example, to select schools by certain predetermined criteria, then to randomly select which of these schools were to be exposed to the programme, and which would be the untreated control group. It would be normal to do a before-and-after study. And it would be normal to try not to let other factors influence some schools and not the others. The United Nations and the World Bank have strict rules on Impact Studies (Clemens & Demombynes, 2010; Angelucci & Di Maro, 2010; Winters, Maffioli, & Salazar, 2011; Winters, Salazar, & Maffioli, 2010). The LEA statisticians volunteered the information that this had not been done in their LEA, and that they had not been consulted in the design of the pilot.

When carrying out an experiment, it is normal to be very careful in selecting the subjects for the experiment. In an experiment such as the Pilot it would be normal to specify the parent population, ‘All schools with less than 40% of pupils reaching Level 4 at Key Stage

4' perhaps, and then randomly select schools from this population, so that the results would be representative of the population.

This was not done. First, the LEAs volunteered: they were not selected from a group of LEAs meeting the criterion of less than 40% of pupils reaching Level 4 (or any other defined group). Nor were they randomly selected. The only group they had in common was, 'All LEAs' and they were self-selected using different criteria – they may perhaps have been LEAs which were persuaded that the ISP would help them be even better, or ones which were desperate to try any new system that might possibly be an improvement on the existing system.

For both the Pilot and the Follow On, individual LEAs used their own criteria for selection of the schools in their own area. These criteria reflected their own priorities, which in turn reflected the social, demographic and economic situation in the individual LEA – and these vary enormously from LEA to LEA.

The Department for Education and Science meant the ISP to be used in schools which had particularly low performance, and the pilot study was meant to cover schools where fewer than half of the pupils achieved Level 4 at Key Stage 2. The LEAs chose completely different schools, samples which were not part of the intended parent population. Nearly half the schools selected for the pilot already a better performance than this. The range was from 25% to 94% achieving Level 4 in 2002 with a mean of 59%. In one LEA, for example, three of the four schools had scores of more than 60%. In another, 10 of 18 schools exceeded the 50% mark, and four had scores of 78% to 93%. In a third, 11 out of 16 schools exceeded the 50% mark. What, then, was the pilot programme examining? It was certainly was not measuring the effect of the programme on the target population.

Similarly, the LEAs selected schools according to their perception of individual headteachers and individual schools. Some chose schools where new headteachers who were perceived to be energetic reformers had just been appointed. Some avoided schools with new headteachers on the grounds that they would be fully occupied settling in. Some avoided schools which were obviously performing very badly but were perceived to be stressed enough without yet another 'intervention'.

Headteachers informed us that they had to deal with 25 to 35 'interventions', usually central government programmes, some of which would require as much staff time and effort as the ISP did. The Pilot did not attempt to control for this, or even to record which 'interventions' were being implemented in each school. The particular mix of 'interventions' varied from school to school, and a different mix applied to teachers at different levels. This raises the question of which 'intervention' was responsible for any change recorded, or, indeed, prevented any change.

Since the Pilot sample was not representative of any parent population, it was not possible to draw any inferences about what it meant for any group of schools except those actually in the Pilot. Certainly it could not be used to draw inferences about the impact on the specific group of schools that we were told it was designed to study. And it was not

possible to do anything resembling an impact study meeting the internationally accepted standards.

Control group

Obviously there must be some control group, or else it could be argued that all similar schools throughout the country had the same change in test scores whether or not ISP was applied, so we could ignore a fall in performance, for instance. It would be normal for a pilot study to identify the schools in the parent population, to pair similar schools, then select which were to receive the treatment and which were to be controls (though this is not a simple procedure). This was not done. There had been no selection of a control group before the ISP so we had to find one ex-post. We decided, for lack of credible alternatives, to use the schools that were selected for the second round as controls. That is to say, we were left with no alternative but to assume that a body of schools had been selected on one set of criteria as needing an intervention to improve performance, and that half of this selection had been selected at random to go into the first phase of ISP, and the rest of them to go into the second phase two years later. Unfortunately this was not how it was done. We know that the selection for the second phase was based on different criteria – the two sets of schools were different even in the years before the Pilot - and we have no idea what bias was introduced.

When we were asked to carry out an evaluation of the second phase, no such control existed as all of the initial selection had been incorporated into the Programme.

Accordingly, it is questionable whether any inferences could be drawn from schools and LEAs selected in this way.

Random variation

The Key Stage test scores for any school varied a lot from year to year, for reasons quite unrelated to ISP. There are purely random elements in any test score. If the same test could be given to identical pupils on different days, we could expect different results. The wording of one year's tests may help some pupils understand, and reduce the score of pupils from some areas and pupils with a different language background. A school may get higher scores in one year than in another because there is a brighter, or otherwise different, group of pupils sitting the tests that year.

The results will be influenced by the competence of the teachers. The results could be changed dramatically because a better or worse teacher was in charge during the year of the test. The team were told that the underperforming schools had a higher level of turnover of teaching staff than other schools, and indeed, there were examples of three or more teachers being in charge of the KS2 class during the test year. There was also a high turnover of head teachers, which would have affected test results. No evidence was produced to support these assertions.

Other variation

A lot of the schools in the ISP had a high level of pupils having English as an Additional Language (EAL), some with as many as 95%. Some of these pupils had arrived as immigrants within a year of two of the test, some were born to recent immigrants, some were second or third generation British, with Polish or Punjabi, say, as the home language of a bilingual family, which is to say very different performances could be expected from schools with the same proportion of EAL pupils. There were no data on which was the case. The team was informed that typically these pupils would perform badly in reading and writing at Key Stage 1. In some LEAs the previous experience was that these pupils made their big improvement in these subjects at secondary level, just after the Key Stage 2 tests. However, if they made the improvement slightly earlier, in primary schools, the schools which had these pupils at both levels of test would appear to be high-performing. If ISP could speed up this improvement, it would mean a sharp improvement in Key Stage 2 results compared with Key Stage 1, but we would have no way of knowing whether it altered Key Stage 3 results. It might have no effect at all.

It cannot be assumed that the pupils doing the Key Stage 1 tests are the same as the ones doing the Key Stage 2 tests four years later, so it cannot be assumed that any change in marks achieved are due to the school. Many of the schools covered had a high level of 'turbulence', with pupils moving from one school to another. This was particularly obvious in inner-London schools. The KS2 pupils might have been in different schools in the same LEA, from different LEAs or from different countries in the KS1 year. It was not possible, to compare the performance of children who were in different countries or different LEAs in the Key Stage 1 year. Comparisons could have been made of performance of children who were in the same LEA for both tests, but this would have required that the ISP project had asked the LEAs to collect the information, which they had not done. In some schools, nearly all the pupils who did Key Stage 2 in one school had also done Key Stage 1 in that school. In one school, just a single pupil in the Key Stage 2 class had been at the same school for Key Stage 1. One headteacher said that the school had a high turnover, as the upwardly mobile families moved pupils to other areas – producing a bias, as the pupils from these families could have been expected to perform better.

Each pupil does have a unique individual identification number in the Department for Education and Science database, and it would, in theory, be possible to link the performance of those pupils who had done both tests in England (but not Scotland or Wales). In practice it would have required far more time and money than we had to make use of this. Individual performances are sensitive information, and in our view it would have been both unethical and unlawful for us to have access to this information for this particular study. Even if we had had this information we would not have been able to determine which of the schools that a pupil had been to should be credited with any improvement. Similarly, we did not have the resources needed to determine which students in a class, if any, benefited.

Some of the schools had particularly severe turbulence, with changes of headteacher, teachers and pupils. In these cases it may be questioned whether the 'school' of one year is in any sense the 'school' of four years later. One bias would be introduced if we ignored these schools, another if we included them. However nearly all of the schools selected had some of these problems. It may well be that schools in richer areas with stable, non-immigrant populations, are more likely to retain their staff, and get better results, regardless of the quality of teaching.

At best, therefore, the figures for Key Stage 1 would give some idea about the performance in the year before the test with the obvious limitation that the pupils in the poor areas covered and those with more EAL pupils might be expected to have low scores anyway. The figures for Key Stage 2 might indicate that the ISP intervention in the final year before the test influenced, or did not influence, results. They also indicate, to an unknown extent, the influence of teaching in the previous years, mainly in the same school for some schools, mainly in other schools for others.

Number of years

We would have liked to analyse year-to-year variations in results over a long time period before the Pilot, so we could make a judgement on the degree to which any change observed might be due to these exogenous factors rather than to the Pilot. Unfortunately the only data series was recent. For Phase 1 we had to do the analysis on only 80% of the schools, those for which 5 years' results were available. Some schools had closed, some had started more recently, and there were problems with the data for others. Accordingly the main analysis was done on those schools for which data from 1998-2004 existed. 115 schools met this criterion out of 143.

Value Added

One figure starting to be produced by the Department for Education and Science when we did our evaluation was the 'Value Added' statistic, designed to measure how much of any improvement, or lack of improvement in pupils' performance could be ascribed to the school. This was comparing Key Stage 1 and Key Stage 2 test results, and was produced by a complex formula using several data series. Certain observations in the data were excluded from the calculation: the Department for Education and Science ignored results:

> 'if at either Key Stage 1 or Key Stage 2 results for that pupil were missing, or the pupil was not eligible for the tests, would take the test in the future, was working at the level of the tests but 'unable to access them', was 'not awarded a test level', was working towards level 1, or if there were lost scripts, or the results were 'annulled', 'disapplied', or 'disregarded'.

It may be that some of these reasons applied for just one of the tests taken in any year. It appears likely that many of the pupils at some of the pilot schools would have been omitted on these grounds, introducing a bias.

The Value Added statistic had all the data problems discussed above, and one discussed below. It required complex calculations, and used explicit and implicit assumptions which necessarily introduced biases. It is normally considered good statistical practice that statistics should be presented with the minimum amount of processing of raw data, and that routine publication of statistics derived by amalgamation of data from different data sets, and complex calculations are to be avoided, to avoid such problems.

The way in which the statistic was calculated might have had some value if we had been comparing middle class schools with static staff and pupils, but was of no value at all for our evaluation.

Experimental Results

The ISP management had not intended to do an evaluation, even though this was a pilot of a new programme of intervention which they intended to roll out to cover a large proportion of the educational system. They did agree to an evaluation because pressure had been brought to bear on them by LEAs which thought that the ISP was badly managed and ineffective. Had the management thought about an evaluation at the beginning, and particularly if they had examined the data needed, a much fuller evaluation would have been possible.

We had discussed data availability with the clients before entering the contract. They assured us that the data were readily available from the Department for Education and Science statisticians within the building, and that we would not need the time we had set aside for data collection, a time we had based on our experience elsewhere. In fact the statisticians within the building flatly refused to give us the data for unstated reasons, and we had to spend a lot of time and money finding out where else we could get the data, which turned out to be in another section of the same Department - where we were told that that the man we first approached was the person who should have supplied it.

We were hampered by the fact that no preparations had been made for an evaluation when the ISP was started. Indeed we were told by consultants that it was only at their insistence half way through Phase 1 that an evaluation was carried out at all. This meant that there was not a properly selected control group. It also meant that a lot of information that was available within some LEAs, like a comparison of Key Stage 1 and Key Stage 2 results, and which could have easily been collected by other LEAs on request – the LEA statisticians we met were keen, helpful and enthusiastic – was not available within our time and budget constraints. Similarly, raw data and processed data that would have been provided by the Department for Education and Science for routine evaluation by one of its own sections was not made available to an outside organization, a university, requiring the information at short notice. The statisticians said that we, as an outside organization, would have been charged if they had produced anything but the raw data at our request. Had we

been given correct information on the availability of data by the clients, we might have done something about this.

Some data, notably matched figures, matching Key Stage 1 and 4 can be bought. We were not sure how well this would allow for the extreme turbulence of the selected schools. It was not possible to buy data, given the financial limits to the study which had been imposed by the clients.

Percentage reaching level 4

The Department for Education and Science targets were based on the percentage of pupils reaching level 4 at Key Stage 2. This is very different from a figure for the average score in a class.

The English government had adopted a policy of monitoring the efficiency of government and local governments organizations by setting 'targets', usually a calculated statistical measure. Sanctions were imposed on management if these 'targets' were not met. There was a widespread view, among the general public at least, that the organizations were skewing their performance to do whatever would produce the statistical measures to show that they had met their target and were forgetting what their real objective was. The health service could achieve their targets by manipulating waiting lists, and neglecting some patients for instance. We considered therefore whether the Department for Education and Science targets could be skewing the delivery of education in this case.

The Department for Education and Science set achievement targets for all schools on a variety of targets. The targets for primary schools were set in terms of the number of pupils reaching Level 4 at Key Stage 2, for example. This target ignores the number of pupils who reach Level 5, as well as those who were working below the level of the test, those who were not awarded a test level at all, and those reaching Levels 2 or 3.

This contrasts with an average score for the school, which takes all pupils into account.

In our interviews, he team were told by consultants and headteachers that the teachers and schools were under strong pressure from the Department for Education and Science and the LEA to reach the target by increasing the proportion getting of pupils reaching Level 4. The teachers were under personal pressure, both out of loyalty to the school and for career reasons. They explained that they could achieve this in several unacceptable ways that did not improve average score and might well reduce it. They said that lots of teachers and schools did in fact adopt these practices, though the teachers, consultants and headteachers we spoke to said that they themselves resisted this pressure and did not use such improper practices.[3]

[3] We did not challenge this claim, of course. Similarly, when I am interviewing the village moneylender and he says, 'I only charge 1% interest, and the farmers never pay me back anyway. I am losing money out of it, and only doing it out of charity,' I nod sympathetically, and say, 'That must be difficult for you.'

The first way was to concentrate on two or three pupils who would otherwise get marks just below Level 4. If there were 15 rather than 12 pupils in a class of 25 reaching Level 4, the target figure, 'Percentage achieving Level 4' would rise from 48% to 60%. This looks very impressive, though it may just mean that three pupils get 5% more marks. How do teachers get the extra time to push these pupils? The obvious way is to ignore the pupils who would normally get a Level 5, and will certainly get a Level 4 anyway, and to ignore those who could never be pushed as high as Level 4. However low their score falls, they did not affect the target statistic.

The tables and graphs below show how this works using a simplified example. By concentrating efforts on the pupils achieving just below Level 4, the percentage achieving Level 4 or above can be raised 27 per cent from 44% to 56%, but this means reducing the average marks 11% from 5.4 to 4.8.

He then goes on, 'But what some moneylenders do is. . . .' And then tells me a tale of violent extortion. I get the information I need on how the system really works, and he saves his face.

COMPARING AVERAGE MARKS AND % LEVEL 4 OR ABOVE

Example 1

LEVEL	Level			Level		Level			Level		Total	
	0 – 2			3	3+	4	4+	4++	5	5+		
Marks	1	2	3	4	5	6	7	8	9	10		
Number of pupils	1	1	2	4	6	4	3	2	1	1	25	
Total marks	1	2	6	16	30	24	21	16	9	10	135	marks
Average marks											5.4	marks
% Level 4 or above											44	%

COMPARING AVERAGE MARKS AND % LEVEL 4 OR ABOVE

Example 2

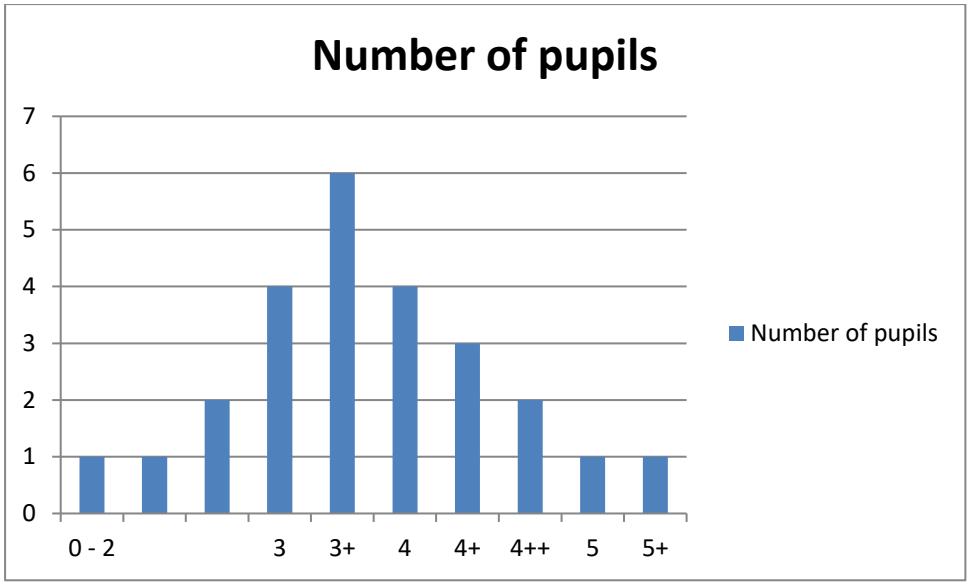
LEVEL	Level			Level		Level			Level		Total	
	0 - 2			3	3+	4	4+	4++	5	5+		
Marks	1	2	3	4	5	6	7	8	9	10		
Number of pupils	2	1	1	4	4	6	4	1	2	0	25	
Total marks	2	2	3	16	20	36	28	8	18	0	133	marks
Average marks											5.3	marks
% Level 4 or above											52	%

COMPARING AVERAGE MARKS AND % LEVEL 4 OR ABOVE

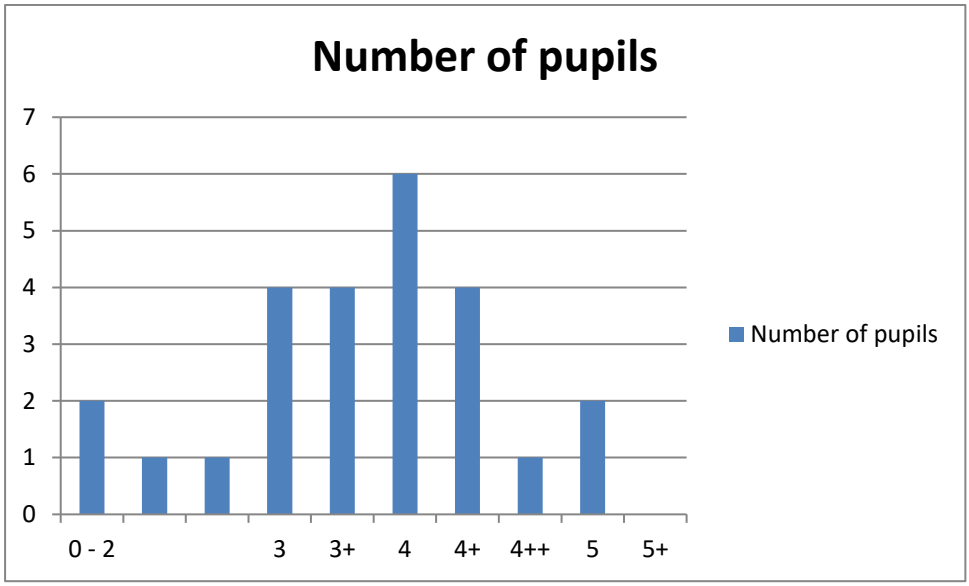
Example 3

LEVEL	Level			Level		Level			Level		Total	
	0 - 2			3	3+	4	4+	4++	5	5+		
Marks	1	2	3	4	5	6	7	8	9	10		
Number of pupils	4	0	0	7	0	12	0	1	1	0	25	
Total marks	4	0	0	28	0	72	0	8	9	0	121	marks
Average marks											4.8	marks
% Level 4 or above											56	%

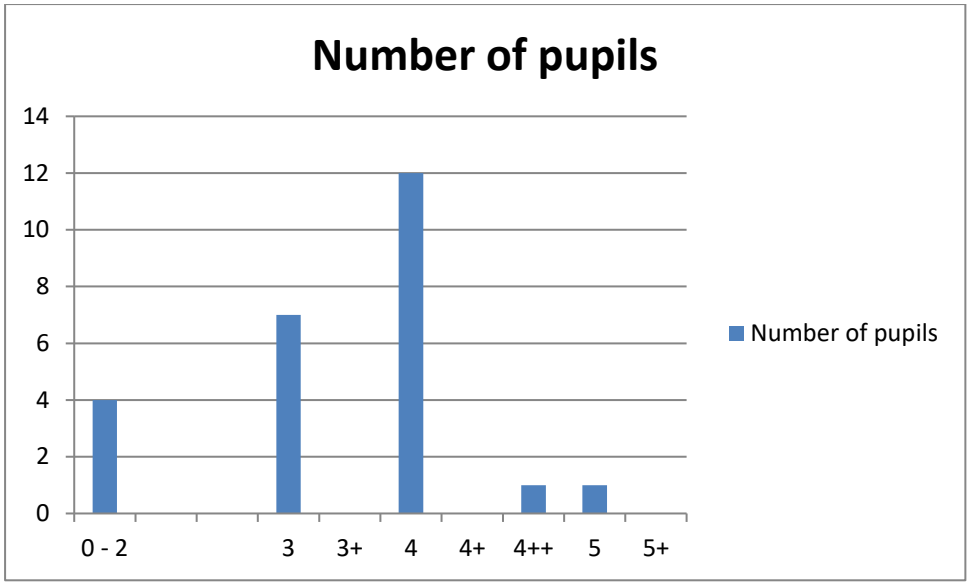
Example 1



Example 2



Example 3



The ISP gave the teachers computer programmes which made this a lot easier: these could identify the pupils who were just below Level 4, so they could be given an extra push. They also helped identify those pupils who were never likely to receive Level 4, and those who might drop from Level 5 to Level 4, but were most unlikely to drop below Level 4.

An alternative way of increasing the percentage achieving Level 4 was to see that some of the lower performers were excluded from the calculation, and the ISP system gave teachers the tools to identify the pupils who should be excluded. If two weak performers in a class of 25 are excluded, the percentage achieving Level 4 might increase from 48% to 52%, meaning that the school reaches the government target figure. One way reported in the interviews was to suggest to parents that a weak pupil might take a day off when the test was taking place, another was to encourage, or permit, parents to take weak pupils on holiday or go on a long visit to Poland or Pakistan to 're-engage with their culture'. Disruptive pupils might be excluded. The Department for Education and Science also reported omitting results when results for that pupil were missing, or when the pupil was not eligible for the tests, would take the test in the future, was working at the level of the tests but 'unable to access them', was 'not awarded a test level', was working towards level 1, if there were lost scripts, or the results were 'annulled', 'disapplied', or 'disregarded', which provide lots of possibilities.

Another, common, intervention was to put the best teachers in the class that was sitting the test and give them additional resources, using the worst teachers in other years. If a lot of the pupils moved from school to school, the school's apparent performance would be maximized by this. Using resources on pupils who would be moving to other schools for their KS4 would be helping the other schools.

Percentage of a percentage

There is a presentation bias that arises from expressing change as a percentage of a percentage. At first sight it may seem that if the percentage achieving Level 4 rises from 50% to 60% there is a 20% improvement in 'performance'. When one tries to work out what it means, one becomes increasingly confused. If a school increases the number of pupils reaching Level 4 by two, the result can be presented as a percentage of a percentage. In one school this may mean that 12 pupils reach the target instead of 10, 48% instead of 40% presented as a 20% increase in achievement. In another it may mean that 25 instead of 23 do, 100% instead of 92%, an increase of 9%. Surprisingly, some of the schools selected did have as many as 96% of the pupils reaching Level 4, which meant that it would be impossible to have 2 pupils increase to this level.

Statements of an increase in average mark are more comprehensible.

Our Conclusions

In Phase 1 it was difficult to identify any impact from the ISP for Key Stage 1. The Average Point Score results were broadly the same before and after the ISP started and broadly the same as for the control group. Average Point Score was not available for Key Stage 2, so we had to use the suspect measure of percent reaching Level 4. There was a sharp rise in the levels reported after the ISP, which was not observed in the control. We pointed out in the report that it was possible to make these increases look much larger or smaller by selecting different base years for the comparison.

Did they concentrate efforts on half a dozen borderline pupils?

In Phase 1 we reported that for KS1, where it was possible to use Average Point Score, rather than a break point, there was no change in performance. These statistics were not made available for KS2, so the possibility remained that borderline pupils were pushed, at the expense of the very good and very poor students. This possibility was taken very seriously by the team and emphasized in our report and in other communications with the clients.

In Phase 2 we had a longer data series and were able to compare the average mark obtained in the Key Stage 2 tests with the percentage of pupils obtaining Level 4. For English there was a clear indication that, overall, average marks fell by perhaps three percentage points over a time that the percentage attaining Level 4 rose 9 points. This discrepancy occurred for all LEAs except one and was much more marked for some of the LEAs. There were only half a dozen schools for which Average Mark increased more than Percentage Level 4. For Mathematics there was an increase in overall Average Marks, of about 3%, compared with an increase in 10 percentage points in the Percentage of Pupils achieving Level 4. This does not show beyond doubt whether teachers were concentrating on the pupils who might achieve Level 4, but is consistent with it. The possibility was considered that the vast majority of pupils are concentrated very close to the minimum Level 4 mark, so

that if all pupils got 3 more marks, this would push 3 or 4 pupils per class into Level 4, but disaggregation of the statistics showed that this was not the case. Disaggregation showed that for a third to a half of the schools, there was there was clear indication that Average Marks fell while the percentage of pupils achieving Level 4 rose. Again, the concentration could be on the pupils who are borderline Level 4, to the detriment of the others

Caveats

To summarize, the only statistics available which could give any indication of whether the ISP had any effect on pupils were the results of the tests. These probably were correlated with what the pupils were able to do, but we had no information on this: the results may well have reflected other factors. There were certainly random and non-random fluctuations on what test result would be produced from a group of pupils of a given ability, and these could be expected to influence different LEAs differently, but again we had no information on this. There were large year to year fluctuations in the average results and percentage reaching level 4 in total, more obviously in each LEA and most obviously in each school. Many reasons could be suggested. The very short time series meant that it was not possible to analyze the fluctuations in an attempt to quantify the random and non-random fluctuations. That is to say any variation in test scores might be just normal variation, which happened to affect the chosen schools in one way. Since these schools were not systematically selected, non-random fluctuations affecting this particular group may well have been due to their special characteristics (e.g. non-English speaking pupils, pupils from poor families, schools with a high proportion of children having special needs, regional culture and dialect). We must be extremely cautious about drawing inferences from them.

The schools were not randomly selected, or selected according to common criteria. There was no attempt at a controlled experiment to evaluate the ISP. The figures were consistent with a large number of schools, sometimes most of them, getting worse results after the ISP.

The figures were consistent with schools adjusting their teaching and other factors to maximize the percentage of pupils achieving Level 4 at Key Stage 2 by reducing the attention given to non-marginal pupils. The percentage achieving Level 4 rose while average score fell, sometimes overall, but always for a large number of LEAs and schools. It is noted that it is possible to get some increase in the percentage Level 4 in a school without reducing the average by these methods, so they do not always show up.

For all these reasons we could give only the most tentative statistical conclusions, except to state that the figures were consistent with concentrating on the marginal pupils at Level 4.

Inferences drawn

Inevitably the inferences drawn purely from analysis of the data will be quite different from those drawn by people who are reading the full report and integrating it with their knowledge, experience and emotion. And people may be unable or unwilling to understand what the data show and do not show, and the limitations in the reliability of any results of a statistical study. However, it is the inferences that are drawn that influence future study, so we must take into account inferences drawn

By team as a whole

By clients

- Department for Education and Science
- National Literacy and Numeracy Strategies
- ISP management team

By consultants delivering ISP

By LEA officials

By headteachers

By teachers

We can only speculate about how these groups interpreted the evidence and conclusions we provided. Obviously there was a very complicated micro political environment within the Department and within each LEA. In carrying out the study and presenting the results and conclusions, we had to bear in mind how they could possibly interpret what we said, and how they might react to it.

We did know from interviews that, among the people from National Literacy and Numeracy Strategies and the ISP to whom we were reporting, there were people who thought that ISP was so self-evidently going to work that it was unnecessary to do an experiment or do an evaluation. Nearly everyone in this group appeared to believe that the pilot stage had been such a clear success that there could be no doubt that it would be rolled out to the whole country. (Or at least anyone who dissented felt it wise to keep their opinions to themselves.) And the message we got was that nearly everyone believed that any evidence that challenged this belief must be invalid.

Controlling the discourse

Could we, the team, or I, the numbers man, control the discourse by the Clients, Department for Education and Science, the National Literacy and Numeracy Strategies section or the ISP management team? Could we control what they told themselves or what they told other sections, other Departments or their political masters?

Communicating statistics

In most of my consultancy the technical staff in my client organization are statistically literate, and people from other organizations who read my report are also statistically literate. I expect them to understand any caveats and limitations stated and not to use my figures where they do not apply. I expect them to understand the figures I give, and to correct politicians if they misunderstand. In this particular evaluation I worked with team members who were numerate academics accustomed to rigorous analysis, though they were not statistically trained – one had a first degree in physics, the other in mathematics. They understood what I was saying. I had assumed that the people in the clients unit, who were graduates - many of them with experience in teaching mathematics – would be able to understand what I wrote. I did not realize that the staff of the clients' unit were statistically illiterate, and many of them numerate only in the sense that they could teach primary school mathematics. Had I realized this I would have presented the quantitative analysis very differently.

Suppression of unpalatable results

Choice of sample LEAs

A key question was whether the ISP was successful in the experimental schools, and whether ISP should be rolled out on a large scale, first to similar schools in more than 360 LEAs in England, then to other schools.

Our team included two education specialists, part of whose job was to interview ISP staff, consultants, headteachers and teachers. Obviously, they wanted to see successful and unsuccessful parts of the ISP, to see what could go wrong, and to find out whether the fact that some were perceived successful was purely down to chance, or was due to unstated criteria, for which no hard evidence might exist, for instance. The clients were strongly opposed to this, and instructed us to do our interviews on only three LEAs. Two of these were ones where the ISP was perceived to be highly successful. One was an LEA where there had been a lot of problems, so many problems that the LEA had demanded an evaluation. The ISP refused us access to other LEAs. In particular, the ISP officials were most reluctant to let us visit LEAs where problems had arisen, or to visit the schools we wanted to visit.

In our view this selection of LEAs to visit was worse than useless when the evaluation was to see whether there was any point in rolling out the ISP to a large number of schools in all LEAs using normal consultants and consultants whose particular charisma might not have appealed to LEA staff.

The LEAs in the ISP had been self-selected as keen to take part, and we could not assume that the results would be in any way similar to all LEAs and particularly to those where schools were in most need of improvement. A wide range of biasing selection criteria for schools has been identified above.

The inclusion of the most schools and LEAs in which the ISP appeared to be outstandingly successful was clearly a waste of time. The ISP management were of the view that the success was because these had a very competent, charismatic, consultant working closely with an enthusiastic LEA which had chosen the schools very carefully to have headteachers who would make the programme work. If the system were rolled out to all underperforming LEAs in that met the target criteria, there would clearly be some equally successful examples, but the proportion that could be expected was much lower than in the experiment. What we wanted to know was what problems had occurred in other schools, whether these problems had been solved, or could be avoided in future, and whether similar schools could be expected to do better in the roll out. We were not allowed to address this problem. We had also carried out our mail survey to find out what parts of the programme headteachers and consultants were not happy with, or were extremely unhappy with. These results could only be of interest if we followed them up with depth interviews to find out why they were unhappy – and the fact that half were not happy with a part of the programme is serious. We were not allowed to do this.

We protested very strongly, to no avail. Eventually the team interviewed some people involved in other areas by telephone. This gave useful results, but the bias was obvious.

All three members of the team were involved in this phase. The other two members of the evaluating team were fully involved in formulating statistical hypotheses and they took the results into account when interviewing ISP staff, consultants, headteachers and teachers.

On the information the other two team members got, they concluded that the ISP was widely popular and seemed to be effective at improving working practices. They confirmed the worry about teachers concentrating their efforts on marginal pupils. They made a lot of recommendations for tackling problems that had been identified. Overall, their report was favourable. They could not comment, of course, on whether ISP had any impact on pupils.

If the statistical analysis of the results and the discussion of problems that had been identified could be suppressed, and removed from our report, therefore, the survey results and the comments could be presented by ISP staff and civil servants as providing overwhelming support for the ISP.

Recommendations

The purpose of a pilot study is, of course, primarily to find out what the problems are. It is not a criticism of staff or, necessarily, of planning that problems are found, and we

do not want to put pressure on staff to conceal problems. The pilot is done precisely to find out if the new system is fundamentally flawed, or if there are problems that need fixing – and there are always problems. The normal approach in evaluation is to exercise tact, by making recommendations for future action to avoid these problems, which appears more positive than a long list of the problems identified.

The ISP management insisted that we did not include the recommendation section in the evaluation reports. 'This should be treated as internal advice to the ISP team who can use them to inform their discussions, if appropriate. The research team were not asked to make recommendations in their contract and so any recommendations are their private advice to the PNS.' 'Move or amend the recommendations of the research team (as this is not a research finding. Separate the finding from the recommendation)', 'Move final bullet point to recommendations. Not a finding.' I found this argument extraordinary. It is quite normal for the major part of a report to address points not specified in the terms of reference, to the extent that the points in the terms of reference are dealt with very skimpily. The points specified in the terms of reference are the clients' first guess at what issues that will be of key importance to address the main objective. They often turn out to be relatively unimportant. The key issues were often not mentioned because nobody knew that they were the key issues: if they had been identified the clients would probably have dealt with them instead of calling in the consultant. That is why people employ outside consultants.

We agreed to suppress the recommendations, though with serious misgivings. I believe it to have been a mistake.

Literature review

Again, the ISP asked us to remove the literature review. 'It is not necessary to make the literature review (currently Appendix 1) public. This was commissioned and paid for separately from the main evaluation and it is therefore justifiable that it should be treated as a separate document for internal use only.' The logic of this statement escapes me.

Average marks

The question of using percentage achieving Level 4 instead of the average marks really upset the ISP project management. In the first phase, we agreed, with misgivings to let this finding be tucked away in the back of the report as long as they undertook to implement our recommendation on it fully and immediately. This was a mistake. In the second phase they claimed, 'The focus on average marks is not one that was explored in the original evaluation and it wasn't part of the brief for the follow up.' Analysing the statistics on changed test performance was, of course part of the brief. Even if it had not, it would be

considered deeply dishonest for an evaluator to suppress important results because they were not specified in detail in the brief.

In the second phase we set out the facts and figures. We met with strong resistance.

> 'There could be other reasons for the fall in average marks. . . . Our concern is that one link is being made and repeated for which there is no evidence.' 'However the data does not provide evidence to show that the fall in average marks is consistent with teachers concentrating all efforts on pupils who could, with a push, achieve Level 4 (p29, 30), or that the concentration could be on pupils who are borderline Level 4, to the detriment of others (p31).'

They did not respond to our request that they present the other reasons that they believed could explain the facts, nor their reasons for the extraordinary belief that the reasons we suggested were not consistent with the facts. This would, of course have required some serious analysis by someone who could at least calculate an average.

Changing the statistical conclusions

The ISP proposed to publish only the main body of the report on the Department for Education and Science, leaving off the appendices, though it costs nothing to put the appendices on a web site.

They proposed to change the executive summary of the findings to

> 'indicate that the statement relating to KS1 Is based on only one years data and a footnote is added indicating that the two year data is now available and shows increases at KS1 also.' 'Any public requests for the report are accompanied by a page of statistics prepared by the [ISP] showing the relevant statistical data for the [ISP] schools (or LEA).'

Clearly it would be totally unacceptable to adjust my conclusions on the basis of calculations by the ISP. I had not seen the new figures, nor the calculations, and I had and have no confidence in the statistical competence of anyone I met working for the ISP.

Providing different statistics

The ISP management told us, 'We intend to publish the main body of the report on the public DfES website; the full report would be available on request from PNS or from NTU. In order to publish it on the website the following recommendations are made: . . .

> 'Any public requests for the report are accompanied by a page of statistics prepared by the ISP showing the relevant statistical data for the ISP schools (or LEA).'

That is to say, our statistical analysis was suppressed, and replaced with statistics produced by someone else, statistics which we were not allowed to see, based on data not given to us, with calculations that were not disclosed.

I do not know who did what to the statistics, but some time later I was informed that recommendations to roll out the ISP on a very large scale had been submitted to the Prime Minister based on the claim that we had shown that the ISP had increased pupils' performance by 14%. The claim of increased pupils' performance was false at many different levels, as shown in this paper.

ETHICAL IMPLICATIONS OF DIFFERENT INFERENCES

We believed that our primary duty was to the children being educated. If an ineffective approach to teaching was introduced, or one which actually reduced attainment, the children would be harmed. Possibly more important, the time, effort and money switched to this approach would be removed from other approaches which might be more effective. If there were weaknesses in this approach, it was our duty to say so in order that they might be remedied. We did not believe that we had any obligation whatsoever to protect the interests of the Department for Education and Science, of the politicians in charge of it, of the National Literary Strategy, or of the ISP, nor did we have any obligation to protect the interests or careers or any of their staff. We recognize from our experience of working in institutions and in government that this perception of duty is not always that of people affected.

The managers who set up and operated the ISP were the people who recruited us to do the evaluation and they kept tight control on what we did and what we published. Their careers would be affected by our conclusions – some were employed specifically for the ISP and their jobs would go if it were terminated. This is a serious moral hazard. UK Treasury guidelines are that the recruitment of evaluators for a project must be done at arms length when the results may adversely affect the recruiters, so the people being evaluated and the people affected by the evaluation can have no control who is appointed to do the evaluation, of how the evaluation is carried out or what results are published. These guidelines appear to be ignored in practice (Bowbrick, 2012). The United Nations agencies have strict rules for evaluation which would not permit this to happen: they have their own pressures on evaluators but the pressures work differently and come from a different direction, usually from the desk officers in headquarters who approved the funding for a project rather than the people who implemented it (Griffiths P. , 2003).

The evaluation team felt itself under pressure. There was also the feeling that we had two different evaluation objectives, summative, whether the project was working, and informative, where we would be helping identify problems and opportunities for future implementation. We have written on this pressure in this case and in others (Bowbrick, 2012; Griffiths, Cotton, & Bowbrick, 2006)

One has, of course, to adapt to the culture of the country in which one is working. International consultants describe with astonishment the British system in which government departments usually send a report back five times for 'corrections', and when the government tender documents say that this should be allowed for in the costings. The UN, World Bank and European Commission have strict rules, saying that the evaluator is employed as an independent, to give an independent opinion. I have evaluated for these organizations, I have been evaluated by them, and I have seen many other evaluations done for them. There has never been any suggestion that the people or organization evaluated could demand changes in the evaluation. If they disagreed with the evaluation, the most they could do was write a response.

The power structure is important here as it introduces a serious moral hazard. Academics in UK education faculties are most likely to get consultancies and research grants from government. There is a widespread belief that if they annoy a government department in any of these consultancies or research projects, they will not get any more. I have been blacklisted for a government consultancy on the basis of a journal article and report written twenty years previously when I was working in another country, though I had an international reputation in the subject (Bowbrick, 2012). While one person may be willing to resist the pressure and lose government funding, there is also a widespread belief that it is not a single person that is blacklisted, but a whole department. There is pressure to submit, to protect the jobs of one's colleagues. One British university had a formal committee to vet all research, not just government research, and to suppress any that produced results that might be unpalatable to any firm or organization that might possibly give the university a consultancy or a grant in future.

WORKING IN AN IMPERFECT WORLD

In this chapter, I have described some of the problems that arose in a fairly normal use of statistics in consultancy research. In this evaluation, as in all research, we had to operate in an imperfect world. Inferences were going to be drawn by someone whether or not we wrote our report, and whether or not it was accepted by the clients, so we had to do the best we could, making unpalatable compromises, which may or may not have been justified. We had to work, as always, with constraints of money and time available. As always the data and statistics were not right for the task. As always the information and support promised by the clients were not made available.As always, the results were not what the clients had expected.

Bibliography

Angelucci, M., & Di Maro, V. (2010). *Project Evaluation and Spillover Effects. Impact Evaluation Guidelines,.* Washington, DC:: Strategy Development Division, Technical Notes No. IDB-TN-136 (Inter-American Development Bank). Accessed at http://www-personal.umich.edu/;.

Bowbrick, P. (1988). Are price reporting systems of any use? *British Food Journal.* , 90(2) 65-69.

Bowbrick, P. (2012). From Economic Research to Policy in 32 Years. *Eurochoices* , 11 (3) 44-47.

Clemens, M. A., & Demombynes, G. (2010). *When Does Rigorous Impact Evaluation Make a Difference? The Case of the Millennium Villages.* World Bank Policy Research Working Paper 5477,.

Griffiths, M., Cotton, T., & Bowbrick, P. (2006). Educational researchers doing research on educational policy: Heroes, puppets, partners, or…? *British Educational Research Association Annual Conference,* .

Griffiths, P. (2003). *The Economist's Tale: a consultant encounters hunger and the World Bank.* London and New York: Zed Books.

Milne, W. (1949). *Numerical Calculus .* Princeton.

Morgenstern, O. (1963). *On the accuracy of economic observations.* Princeton N.Y. : Princeton University Press,.

Winters, P., Maffioli, A., & Salazar, L. (2011). Introduction to the Special Feature: Evaluating the Impact of Agricultural Projects in Developing Countries. *Journal of Agricultural Economics* , Vol. 62, No. 2, 393–402 doi: 10.1111/j.1477-9552.2011.00296.x;.

Winters, P., Salazar, L., & Maffioli, A. (2010). *Designing impact evaluations for agricultural projects: Impact Evaluation Guidelines.* Washington DC: Strategy Development Division, Technical Notes No. IDB-TN-198. (Washington, DC: Inter-American Development Bank, 2010). Accessed at http://www.iadb.org/document.cfm?id=35529432;.