Pitfalls in Educational Research

Published in BERA Bulletin, Autumn 1976, Vol. 3, no. 1

Jack Wrigley, ***University of Reading***

1. Introduction

The British Educational Research Association is very young but one tradition that is already established is that the incoming President should "do his own thing" in his lecture to the conference. So my topic is not designed specially to fit into the theme of our other discussions this week. Rather, it is meant to allow me to talk within my current interests, drawing on my experiences over the years. I shall, however, try to take into account the intended interdisciplinary nature of B.E.R.A., though I fear that my own bias towards empirical and statistical methods will be only too obvious. My topic was chosen nearly a year ago in the first wave of enthusiasm after leaving the Schools Council. I wanted to leave aside, temporarily at least, the problems of curriculum and examinations, and I knew that in my university teaching I should be dealing with the methodology of educational research. Judging however from the reactions of some of my friends to my choice of topic I think that I may have set a trap for myself. Neville Bennett told me that he would not know how to treat such a topic as "Pitfalls in Educational Research". I found that amusing since it would be easy to plan the whole lecture around his recent research into "Teaching Styles and Pupil Progress". No, do not misunderstand me! I believe that Entwistle and Bennett are two of our most competent members and that they are to be congratulated on their work. But even with research of such quality one could illustrate a number of well-known pitfalls - either real - or imagined by some of their more hostile critics. I wouldn't do it. But I am tempted!

One of my other friends suggested that anyone could make a catalogue of pitfalls implying that your incoming President in "doing his own thing" would be his usual nasty, negative self. Finally, for this introduction, I realise that as I am now Chairman of the Educational Research Board of the Social Science Research Council that my talk, though personal, may be thought to have overtones derived from the knowledge gleaned from considering applications for research grants. Since the rejection rate in E.R.B. is high there must be many pitfalls. Indeed as I mention my friends I wonder if I shall have any left if I stay as Chairman of E.R.B. much longer. Today, however, my intention is to be personal and not to speak on behalf of S.S.R.C. in any way. I may indeed produce a catalogue of faults but I shall try not to be too negative.

2. The Nature of Educational Research

Why is it that educational research has such a bad name among the reasonably informed general public? To the extent that the recent Permanent Secretary at the D.E.S. could

assert that a few minutes thought would bring greater illumination of a problem that many an educational research project. And make no mistake, he spoke, not only for himself, but for others too. The main reason, I believe, is to do with the nature of educational research itself. The most important aspect or phase of any research is centred around the choosing of a problem and its reduction to manageable research. There *is* no point in researching into problems which can be solved by ten minutes thought. In contrast, there are some important problems which are so complicated that they will not yield to research. For the beginning student talking to his supervisor for the first time, for the post-doctoral researcher embarking on his career of unsupervised independent research, for the established researcher seeking money from S.S.R.C., for all, the act of selecting a worthwhile yet tractable problem is crucial. It is quite difficult to avoid the trivialisation which so often comes with the reduction of the initial problem into research terms.

3. Pitfalls of Logic (or Lack of)

Let us now turn to Pitfalls Proper and perhaps you will forgive me if I begin with what I know most about - empirical research and statistical methods. Here pitfalls abound - but the mistakes are as often those of logic as of the inadequate application of statistical techniques. Let me give one or two examples, real and imaginary, of lack of logic.

(1) A researcher gave his subjects samples of (a) whisky and water, (b) brandy and water, (c) rum and water. When his subjects became drunk he concluded that the intoxicating factor was water, the common element. A failure of logic. Had the researcher been more perceptive he *would have found* a common element - alcohol. So was it a failure of logic or technique - or both?

(2) The mixing up of correlation with causation gives rise to endless jokes. My own favourite is an old one - the positive correlation between the number of storks nesting in the rooftops in Bavaria and the birth rate. Of course many such spurious correlations are caused by a common increase in both variables through time. Perhaps though we can give this old chestnut a new twist by noting that the role of the stork in conception is confirmed by the fact that now there are less storks there are also less births. So replication, a process I shall advocate as an antidote to pitfalls is *not* always a sufficient corrective procedure!

(3) These are jokes, certainly. But do not forget that the great British statistician, R.A. Fisher, pointed out that lung cancer might be the cause of smoking. It is true that at the time he made his remark he was in the pay of the British Tobacco Company - but he had a point in strict logic at that time.

(4) Let us take the example of monitoring standards in public examinations, concentrating on the problem of comparability between years. As syllabuses change over the years it is difficult to make direct comparisons so one possible approach is to use a verbal reasoning test (v.r.t.) as a common element to measure the calibre of the candidates from year to year. If, on the average, candidates of equal calibre,

measured by the v.r.t. obtain comparable grades then standards are maintained. If the grades improve for candidates of equal calibre then standards have fallen. But have they? Might not the teaching have improved? Might not the v.r.t. itself have become susceptible to changing styles of teaching. Logic and technique are interwoven into this discussion. Consider the fact that there is a sex difference in favour of girls on v.r.t.s. So from the point of view of calibre it could be argued that standards for girls are higher than those for boys, or that boys are treated more leniently than girls. An artificial example will perhaps make my point more clearly.

Example

	Average Score in Examination		Average Verbal Reasoning Score	
	Girls	Boys	Girls	Boys
Group A	53	53	125	120
Group B	48	48	120	115

In the example girls and boys of equal calibre (v.r.s. = 120) score differently on the examination (the boys average 53 is higher than the girls average 48). So the boys have been treated more leniently.

But as Euclid would have said this is absurd since sex was not a consideration in the marking of the examination scripts. So what of the use of a common element measuring calibre in other situations?

Note that my examples so far are *not* pitfalls in statistics but in logic, pure and simple. Let us now turn to some statistical pitfalls.

4. Pitfalls in Statistical Technique

The most obvious and ubiquitous pitfall, mentioned by all lecturers and writers of textbooks is the failure to plan the experimental design and the consequent statistical analysis at the outset. It is also the most common as any statistician will testify. The world is full of those who ask for statistical advice when it is too late. Equally ubiquitous are those who use the statistical confidence levels at 5% and 1% in inadequate and misleading ways. One of my most treasured memories is of the late Frank Warburton at an M.Ed. oral examination in Manchester. I asked one of his students what was meant by the 5% level of confidence and the student replied that the event (or the difference) would have occurred by chance one in *ten* times. Warburton, with a straight face said to the student "And I suppose that had the difference been significant at the 10% level that would have meant one in hundred times by chance". The student said "Yes". We do not, of course, often meet with such crude errors. But we do get those who mix up educational and statistical significance, and those who seem unaware that a difference

which is not significant at the 5% level but which has a probability value p = 0.1 (i.e. is significant at the 10% level) has odds of 9:1 in favour of it being *not* due to chance. I am afraid that most ordinary mortals are unhappy with the inevitable uncertainty generated by the concept of statistical probability. We would do well in educational research to put less store on our probability levels (which at best only give us permission to proceed to make inferences) and to devote more thought to genuine replication of independent experiments upon common topics. For truly independent experiments the probability multiplies, so two experiments producing results significant say at the 10% level together give a combined probability of 0.1 x 0.1 = 0.01, i.e. significance at the 1% level. But I am making a point not only concerned with statistics. When one takes into consideration the complicated nature of almost any worthwhile piece of educational research it is likely that more replication would prevent the spread of false knowledge derived from uncertain conclusions and add to the stock of accepted findings. As always, it is the constant undetected biases in our work which lead to false conclusions, not the random errors which can be allowed for in the sampling theories and the statistical analysis.

Perhaps the most pervasive pitfall in statistics is the failure to appreciate the effects due to regression. The fully grown sons of tall fathers are on average shorter and the fully grown sons of short fathers are on average taller. Galton called this regression to mediocrity and unlike some people he knew exactly what he was talking about. He knew the effects were simply due to regression and that if one regrouped and considered tall sons their fathers would be on average shorter. If there were no correlation between heights of fathers and sons the regression would be complete. Whenever we classify people into superior and inferior on some attribute their measured performance on another occasion reverts towards the mean, except in the case of perfect correlation. Perhaps I can recall an example provided by my friend and mentor, the late Stephen Wiseman, in one of his lectures years ago. Divide the population into three groups on the basis of their attendance at the cinema - heavy, average and low attendance. Measure the average height of the three groups. There will be no difference. Hence attendance at the cinema narrows the gap in height - with obvious biological utility since the short people grow taller and can then see over the heads of the erstwhile tall! You might consider that example as a simple joke noting that there is no correlation between attendance at the cinema and height so the regression is complete. But a genuine experiment in the 1930s divided children into three groups - Bright, Average and Backward on the basis of measured l.Q., subjected them to an experiment using sound films as the teaching instrument - measured their acquired knowledge in standardised score form - found the inevitable regression - and concluded that sound films had narrowed the gap and was therefore a particularly suitable method for use with backward children. Fallacies due to regression are widespread and can trap the most able researcher. The real difficulty is to unravel the *real* difference from those simply due to regression. Even in the original Galton example the argument is affected by the fact that the nation as a whole is growing taller so the over-all mean shifts to complicate the analysis.

There are a number of pitfalls associated with the technique of factor analysis, though perhaps less than in the old days. Factors should be regarded as principles or elements of classification leading to a parsimonious description of the data. Godfrey Thomson wrote

eloquently on the dangers of reifying factors in "The Factorial Analysis of Human Ability" in 1938. "Even in physical or biological science, the things which are discussed and which appear to have a real existence to the scientist, such as "energy", "electron", "neutron", "gene", are recognised by the really capable experimenter as being only manners of speech, easy ways of putting into comparatively concrete terms what are really very abstract ideas. With the bulk of those studying science there exists always the danger that this may be taken too literally, but this danger does not justify us in ceasing to use such terms. In the same way, if terms like "mental energy" prove to be useful, and can be kept in their proper place, they may be justified by their utility. The danger of "reifying" such terms, or such factors as g.v. etc. is however very great, as anyone realises who reads the dissertations produced in such profusion by senior students using these new factorial methods".

Godfrey Thomson was writing with factor analysis in mind but his remarks can obviously be generalized to encompass many theoretical constructs within the realm of educational research. I would like to register for posterity another remark made by Thomson in private conversation with me a few months ago before he died. He said "Sometimes, Wrigley, I wish that factor analysis had not been invented. It has been so much abused". The basic difficulty at that time was the lack of a unique solution which was psychologically meaningful. The only reputable unique solution (that of principal components) was not easily interpreted, whilst the various possibilities with rotated axes were so numerous as to defy unique interpretation. "Garbage in - Garbage out" – a phrase now used with reference to computer programming was then applied to factor analysis. Computers have improved the situation with regard to factorial analysis - the standard programmes yield solutions which at least would be found by independent researchers and which are often psychologically meaningful. Even so there are pitfalls enough. It is not generally realised that where there is no variance there is no factor. But this does not necessarily mean that there is no important effect at work. If, for example, schooling or teaching produced a uniform improvement by a massive but equal amount for everyone no 'schooling' factor would appear since the factor only measures the variation. Perhaps de-schoolers should think again.

Another statistical technique which should be used with caution is that of partial correlation. Wherever one partials out a variable it is never entirely clear what has been eliminated, especially if the variable to be removed is a wide or varied one. I would prefer to partial out age rather than l.Q. score, size of family rather than scores in an English test. I suspect what I am now saying seems fairly obvious but it is perhaps less clear to you that the same considerations apply to analysis of co-variance. This is one of those seductive techniques now made easy by the availability of standard computer programmes. The underlying idea behind co-variance analysis is exactly the same as partial correlation. One should use the technique with extreme caution and take a long hard look at the variable which is supposed to be held constant.

5. False Dichotomies

I could go on for a long time with more statistical fallacies but some of you will have become very impatient with me at this stage. You may feel that I am not speaking to your condition. Many modern educational researchers do not use a detailed research design, specified in advance, and are reluctant to apply classical statistical procedures. They have taken to heart the point I made earlier that reducing a real live problem to a researchable one trivialises the outcome, and have acted by employing a more open type approach. We can see these methods in operation by the illuminative evaluators, some sociologists, many curriculum developers, action researchers, and experts in classroom. It is as if many researchers have instinctively felt that the pitfalls I have so far outlined might be avoided by a return to more subjective methods and the employment of a variety of observational techniques. E.R.B. certainly now receives many research applications which are more open-ended and less tightly planned than they used to be. In spite of my bias towards an older tradition I welcome this trend, particularly if it leads to an attack on more worthwhile topics. We must however avoid the danger of false dichotomies. Some of our illuminative evaluators are able to operate in open-ended situations in such a way that by skilful use of cross-checks within the research team and sometimes independently from outside they can minimise the subjective element in their work. There is no doubt, however, that the pitfall of undue subjectivity is a real one. The danger of a false dichotomy arises when highfaulting attacks are made on the psychometric model in a highly theoretical way by pointing out that the agricultural model, developed by R.A. Fisher, is not suited to education and should be replaced by an anthropological model. I find such a highly theoretical discussion less than helpful and in some ways dangerous. Methods experiments with trial and control groups fail for technical reasons to do with the complexity of the experimental situation not because analysis of variance techniques were designed for agriculture rather than education. Evaluation is an essential part of most educational research and much of it must inevitably be rough and ready. We should use eclectic methods, objective and subjective, precise or rough as appropriate. If we can use fully designed test or measuring instrument alongside our subjective judgment then the two together should help us make a more informed over-all assessment. The best of the illuminative evaluators are well aware of these points just as the best of the psychometricians have always been aware of the danger of undue narrowness in their investigations.

6. Over-Sophistication

One of the most tempting pitfalls for those like me, interested in statistics, is to use over-sophisticated techniques. Much of the data in educational research is extremely rough and no amount of elaborate treatment of that data can compensate for that roughness. An elaborate analysis of variance or factor analysis will not be valid if either (a) the original data is inaccurate, or (b) the categorisation is false. Wherever possible an observer should be present when data is collected from school situations and spot checks should always be made when questionnaires are used. The computer is to blame for some of our over-sophistication. No, of course not the computer, those who use it! In order not to appear as too much of a Luddite I shall not elaborate my criticism of the use of computers

beyond saying that their uncritical use can lead to (a) over-sophistication beyond both reason and comprehension, (b) lack of insight formerly derived from playing around with the data, (c) lack of real understanding of techniques on the part of researchers using standard programmes which are often unsuitable, (d) delay. The last may seem odd but amazingly analysis sometimes take longer when the computer is used than they would with hand calculations.

Conclusions

My own view is that educational research is an activity akin to engineering - a problem solving subject drawing on the disciplines of anthropology, philosophy, psychology and sociology for its insights and techniques. B.E.R.A. ought to be an excellent organisation to promote the kind of interdisciplinary study necessary for good research. The bringing together of workers from the various disciplines should help to expose the pitfalls due to blinkered thinking. When psychologists can challenge sociologists about their use of social class whilst sociologists attack the concept of general ability it perhaps takes a philosopher to point out that both concepts are curiously alike - useful and pragmatic - no more real than the reified concepts mentioned by Godfrey Thomson - and occasionally inhibiting both thought and action.

This brings me to my final pitfall. Educational research is so difficult and complex, there are so many pitfalls that it is easy to become inhibited by critics and by one's own doubts. It is tempting to give up research for either teaching or administration - to talk rather than do, to opt for the quiet life when one's application to E.R.B. is rejected. We are so short of good active educational researchers that I should be sad if any words of mine about pitfalls should inhibit the researchers of the future. Not only would I encourage them to act boldly and to attack important topics whatever the difficulties, I would also encourage them not to sell themselves short. One pitfall that I am not worried about is the mixing of the opinion of the researcher with his findings from the research itself. I have little patience with those researchers who will give no opinion, letting the research speak for itself. The researcher should separate his opinions from his data, and should make explicit his value judgments. He should not, though, separate himself from the decision makers and the planners. That way is the worst pitfall - to deny the full importance of the findings and the opinions of the educational researcher.