EXPERIMENTS in EDUCATION

Statistical Design and Dynamics

By PROFESSOR N. GOURLAY

Professor of Education, University of the Witwatersrand

This is a summary of Professor Gourlay's paper. Owing to the fact that he used the blackboard and other illustrations, it was impossible to reproduce the paper in toto.

INTRODUCTION

WHEN some new technique or method arrives on the educational scene there are at least three ways of reacting to it—one may welcome it with enthusiasm, one may more or less ignore it, or one may approach it with an open mind but with interest and, stimulated by that interest, seek to evaluate it.

I think it was because the organisers of this Conference felt that some of you might like to study the value of automatic teaching experimentally that I was asked to make this contribution to the Conference proceedings. I should also hope that what I say this afternoon will not only make some of you a little more informed about the principles of experimental design and statistical analysis but, in consequence of this, more able to assess the value of any experimental studies you might meet. I say "some of you" for I am very much aware that many of you in the audience know much more about statistics than I can present in a short session such as this.

I should also like to add that what I shall say is better said in a text-book which I always recommend to my students. It is E. F. Lindquist's *Statistical Analysis in Educational Research*. The book came out in 1940 and is now out of print but I haven't yet seen any book on statistics which is more suitable for students in education, at least from the point of view of understanding fundamental principles.

Like Lindquist, I shall concern myself only with experimentation carried out in schools. The ordinary laboratory experiment is a somewhat easier matter to control. In any case the fundamental principles are the same.

Basic principles

What is basically involved in any educational experiment? Briefly, the following:

(i) *Sampling of pupils.* Two or more samples of pupils or students are required. The samples may be random, matched, etc. (In all cases, some element of randomisation should be involved.)

(ii) *Application of treatments.* Normally each sample receives only one treatment (or method). Thus, in an experiment to evaluate automatic teaching, the simplest arrangement would be to have one sample receive automatic teaching and to have another sample taught by ordinary classroom methods.

(iii) *Criterion test.* After the application of the treatments over some suitable period of time some criterion test or tests must be applied appropriate to the evaluation of the treatments being considered.

(iv) *Statistical analysis.* Some process of statistical analysis must be applied to the criterion test data to yield a conclusion which will be expressed in statistical terms.

The crux of the experiment is the mean difference(s) between the samples or treatment groups on the criterion test. Thus, in the simple case of one sample receiving automatic teaching and another sample receiving ordinary classroom teaching, the mean scores of the two groups on the criterion test might be as follows:

$$M_{Aut} = 15{\cdot}2 \qquad M_{Ord} = 11{\cdot}7$$

The important question is—to what extent is the difference between the mean scores due to a real difference between methods? Alternatively, to what extent can the difference be attributed to other factors which are or might be operating?

These factors include (*a*) pupil differences (sampling error), (*b*) teacher differences or bias (teachers not only vary in ability but may favour one method more than others), (*c*) a school factor or bias (the relative success of different methods may vary from school to school) and (*d*) other factors such as differences in classroom conditions.

The problem of statistical analyses is to find out the probability of attributing the difference between means to these factors only. For obvious reasons these factors are known as error factors (they tend to hide any real difference between methods which may be present).

The statistician solves the problem by obtaining a measure of the error involved against which he compares the difference between the group means. As a result of this comparison (critical ratio, t-ratio, F-ratio, etc.), he finishes up with a probability statement to the effect that there are $\times$ chances in 100 of obtaining a difference as great as the one actually obtained on the assumption that there is no real difference between methods and only the other factors, listed above, are operating.

If $\times$ is small (F or $\times = 1$ and 5, one speaks of the one per cent and five per cent levels of significance) the investigator might be inclined to reject his assumption and instead assume that there is a real difference between methods. Strictly speaking, he never *proves* a real difference although he might show it to be very likely.

The magnitude of the error involved in an experiment determines the so-called *precision* of the experiment.

Obviously, precision is inversely proportional to the amount of error.

$$\text{i.e. precision} \propto \frac{1}{\text{amount of error}}$$

In experimental design one therefore tries to make the errors involved as small as possible so that there is a better balance of "showing up" a real difference present.

Limitations of simple methods experiment

The simple type of experiment already mentioned—two or more groups, usually in the one school, taught by two or more methods—is of very limited value. One can obtain an estimate of the error due to pupil sampling but no estimate of the error due to the other factors (teacher differences, school bias, etc.).

A significant result obtained for such an experiment cannot be regarded as holding outside of the actual school, teachers, etc. employed in the experiment. What is required is a generalised result which will hold for a whole population of schools with their different teachers etc. This requires an estimate of error which covers all the error factors.

Experimental design for generalised result

Without any more ado, I present a diagram which satisfies the necessary requirement:

		Methods (two, three or more)		
		A	B	C
	1			
	2			
Schools	3.	One class in each of the cells of the table		
	.			
	.			
	.			
	.			
	n			

Points to note are:

1. A random sample of schools is taken from the total population of schools.
2. In each school, random or matched samples or even intact classes are assigned at random to the two or more methods involved. (*A*, *B* etc.)
3. All other factors, teachers, classrooms, etc. are randomised.

The story does not end here.

1. If the investigator wishes to study the effect of such factors as age and intelligence on the methods it is very easy, once the basic principles are understood, to make the necessary modification to the experimental design and the statistical analysis.
2. Matched groups give the best precision. Where matching is not possible (sometimes intact classes must be used) near equal precision can be obtained by the use of the *analysis of covariance*. This requires the use of a pre-test, i.e. a test prior to the application of the methods. The pre-test should correlate as highly as possible with the criterion test.

3. Probably not enough attention is paid in experimental work to the criterion test. One method might be better than another on one criterion but on a different criterion a different result might be obtained. For example, automatic teaching might prove better than ordinary classroom teaching on an objective type of test but not on an essay type of test.
4. The time factor can also be important. One method might be better than another as judged by a criterion test given immediately at the end of the experiment. But it is important to know what difference there is after some lapse of time—in other words, how retention depends on the methods. There is therefore the need for a follow-up test.
5. There is another aspect to the time factor. A new Method *A* might be better than an old Method *B* as judged by preliminary experiment. But one wants to know whether this will be maintained. New methods bring with them an initial enthusiasm which might account for better results but this enthusiasm (and therefore the better results) might evaporate with the passage of time. There is therefore the need for a prolonged period of experimentation quite apart from the idea of a follow-up test.

A few final remarks:

(i) The paucity of educational experimentation is remarkable.

(ii) Quite apart from lack of statistical knowledge, there is the practical difficulty of getting authorities and headmasters to agree to re-organisation of classes and instruction for experimental purposes.

(iii) Even when one is able to carry out experiments, the rewards are sometimes dubious. Thus, you might get someone argue that there is little point in generalised results — what is good for one teacher need not be good for another. And sometimes one can make oneself unpopular by the results one obtains from experiments. Enthusiasts for new methods do not like to have their claims for these new methods refuted.