MEASURING THE EFFECTS OF TEACHING ON LEARNING

John Church
Education Department
University of Canterbury

This paper extends the analysis in Church (1992b) to identify the main technical requirements which need to be met during attempts to measure the relative effects of different instructional procedures on learning. These include accurate records of the day-to-day learning rates of individual subjects, controlled experimental manipulation of instructional variables, and experimental designs which demonstrate the reliability of any changes which are observed in rate of learning. These requirements are almost never met by the between-groups designs which continue to be used in learning experiments. The case for within-subjects measures of the effects of teaching on learning is illustrated using a variety of student-administered teaching experiments.

One of the aims of the scientific study of learning is to identify variables which are functionally related to learning. These variables may be variables which determine whether or not learning will occur or variables which affect the rate at which new skills and understandings are acquired by the learner.

The most accurate way of measuring the effects of particular instructional variables on learning is to set up an experiment in which the instructional variable of interest is systematically manipulated while the effects of these manipulations on rate of learning are measured.

Experimental analyses of learning are currently being undertaken using two rather different experimental procedures. These are the between-groups procedure in which the measure of learning is the mean performance change of a group of learners, and the within-subjects procedure in which the measure of learning is change in the performance of individual learners with the passage of time.

The between-groups experiment may be appropriately used to measure the effects of an instructional programme on the mean achievement level of a representative sample from some population of learners but this kind of experiment remains an extremely unsatisfactory way of identifying the instructional variables which affect the rate of learning of individual learners.

Paper presented to the Second Joint AARE/NZARE Joint Conference, Geelong, Australia, November, 1992.

Part 1
Characteristics of the phenomenon being analysed

Learning is said to have occurred when a learner demonstrates that they can now do something which they could not do before or when a learner demonstrates that they can now do something better than they could before.

It is individuals who learn. If there is to be a science of learning, its aim must be to explain and to predict the learning of individuals. If the aim is to explain and to predict learning in individuals, then the search must be for instructional and other kinds of variables which are functionally related to the learning of individuals.

One of the problems with the between-groups paradigm which has dominated the study of learning for much of this century is that it leaves the learning of the individuals in the group unanalysed (Ballard, 1986; Cooper, Heron and Heward, 1987; Johnston and Pennypacker, 1980; White, 1984). "Knowing that the average performance of a group changed in a given way tells little about the performance of individual subjects. It is quite possible that the average performance of subjects in the experimental group improved while the performance of some subjects stayed the same and the performance of others deteriorated" (Cooper, Heron and Heward, 1987, p 232).

Learning involves change. Change cannot be measured by making just a single observation. The most detailed record of change is the one which is produced by frequent and repeated observations of performance while it is changing (Church, 1992b).

One of the major weaknesses of most of the published between-groups analyses of learning is that the performance of the learners has been observed only once or twice during the course of the experiment. When observations are as infrequent as this, it becomes impossible to describe the precise nature of the change which has occurred (White, 1984). A single observation at the beginning and the end of a sequence of instructional interactions is unlikely to identify the particular kinds of instructional interactions which lead to particular kinds of learning.

Only changes in performance can be measured at the present time. Two different languages have developed to describe the changes which occur during learning. The behaviourist is content to observe a change in performance and to talk about this as learning. The cognitivist also observes changes in performance but argues that these are merely manifestations of the changes which have occurred in the mind of the learner. Unfortunately, changes in the mind of the learner cannot be seen or measured at the present time. Given that both the behaviourist and the cognitivist base their conclusions upon observed changes in the performance of the learner, it is appropriate that measures of these changes should remain the dependent variable in learning experiments.

When performance changes it can change along any one of a number of dimensions: topography, accuracy, generalization, speed, fluency, endurance, and so on (Haring and Eaton, 1978; Johnston and Pennypacker, 1980; West, Young and Spooner, 1990; White and Liberty, 1976). These distinctions are important because changes in different dimensions of performance typically require rather different measurement procedures if they are to be accurately recorded.

One of the shortcomings of many between-groups analyses of learning is that the measures of "retention" and "recall" commonly used during such experiments fail to distinguish between improvements in accuracy, improvements in fluency, improvements in generalization, and so on. In some cases these tests fail even to distinguish between skills which the student can and cannot perform prior to the experiment. It seems unlikely that tests which fail to distinguish between what the learner can and cannot do prior to the experiment, or which fail to distinguish between different kinds of learning, could yield results which are found to be reliably associated with particular instructional variables.

Part 2
The technical requirements of productive experimentation

An experimental analysis is one which seeks to identify functional relationships between particular independent variables and the dependent variables which are the objects of the analysis. In the case of a scientific analysis of learning, the dependent variables which are the objects of analysis are certain types of behaviour change and the independent variables are those naturally occurring events, instructional events, and teaching events which can be examined to see if they are functionally related to learning.

In order to identify variables which are functionally related to learning, certain technical requirements must be met. These include (a) the accurate measurement of the learner's level of skill prior to instruction, (b) the accurate measurement of learning during the experiment, (c) the conduct of controlled experimental analyses of the effects of various values of the teaching variable on learning, and (d) the conduct of replication experiments which explore the reliability and generality of the results obtained. The data obtained from such experiments must also be accepted by the scientific community as data of some scientific importance (Sidman, 1960).

Measurement of the pre-experimental skills of the subjects

Experiments involve subjects and the pre-instructional skill level of the learner has a greater effect on the rate of learning of new skills than any instructional variable (Church, 1976). Only if the pre-experimental skills of the learner are assessed with care is it possible to develop a measure of learning which is uncontaminated by the previously acquired skills of the learner (Church, 1992b). An accurate assessment of the pre-experimental skills of the learner is also necessary before the experimenter can describe the kinds of learners to which the results of the experiment are likely to apply. Poor descriptions of the pre-experimental skills of experimental subject(s) is perhaps the most serious weakness in current reports of learning experiments. If the result of a learning experiment is to have any meaning, it follows that the experimental analysis must begin by measuring the pre-experimental skills of the subjects - at least insofar as these are relevant to the new skills which are likely to be learned during the course of the experiment.

The measurement of learning

The second step in the design of any learning experiment is the development of the measure of learning. The most detailed picture of learning is that which is provided by the regular and repeated observation of performance throughout the experiment. The measurement procedures which are developed for a scientific analysis of learning must meet criteria which are acceptable to the scientific community. At the very least, the measurement procedures which are used must be ones which (a) generate an accurate representation of the performance changes which occur during the experiment, which (b) are reproducible so that the experiment can be replicated, which (c) produce results which can be reported in standard units, and which (d) are sensitive to changes in the behavioural dimensions which are being studied (Johnston and Pennypacker, 1980; Simkins, 1969). A more detailed description of the requirements of good measurement (during learning experiments) will be found in Church (1992b).

The measurement of experimental effects

If a learning experiment is to generate data of any scientific value, the experiment must be designed in such a way as to provide an accurate and reliable measure of the effects of the independent variable manipulation on learning. When we ask whether experimental data are reliable, we mean "Will the experiment, if repeated, yield the same results?" (Sidman, 1960, p 43). Note that this question can only be answered if the experiment is actually repeated. This is true regardless of whether the experiment uses a within-subjects or a between-groups procedure.

When a measurement procedure is a new one, a single application of that measurement procedure cannot tell us anything about its reliability. When an experimental analysis is a new one, a single experiment cannot tell us anything about the reliability of the treatment effect which is being reported. The only way of ascertaining whether or not a given independent variable manipulation has a reliable effect on rate of learning is to repeat the experiment under conditions which closely approximate the conditions which operated during the first experiment. This kind of replication is usually referred to as "direct replication" (Sidman, 1960). A probability statement regarding the difference between two means can never be a substitute for direct replication and we mislead ourselves if we continue to believe that it can. It is possible that the individual subjects within a between-groups experiment could provide evidence of experimental reliability, but only if the learning of each subject is affected in a closely similar way. This almost never happens in a between-groups experiment no matter how carefully the subjects are selected (see, for example, Martin and Church, 1992).

The technical requirements which need to be met during experimental analyses of performance change in individuals have been described by a number of writers (e.g. Sidman, 1960; Simkins, 1969; Johnston and Pennypacker, 1980; Barlow and Hersen, 1984; Cooper, Heron and Heward, 1987).

First, the measure of experimental effect needs to be a within-subject measure. In a between-groups analysis the most powerful variable of all, the prior learning history of each subject, is always left uncontrolled. Only the baseline logic of the within-subject experiment controls for differences in the prior learning history of the experimental subjects. Secondly, the experimenter must devise a reproducible procedure for manipulating the independent variable so that this experimental manipulation can be reproduced by the experimenter during the direct

replication phase and reproduced by other experimenters if there is any debate regarding the reliability of the experimental findings.

Thirdly, the experimenter must observe, record, and report the values of the independent variable which operated during the several phases of the experiment so that the independent variable - dependent variable relationship can be quantified and reported.

Fourthly, the experimenter must devise procedures which provide reasonable levels of experimental control over both the independent variable manipulation and those extraneous variables which might conceivably operate to provide alternative explanations for any experimental effect which is observed.

Fifthly, each new experiment should be replicated. This is most commonly achieved by designing the experiment as an ABAB experiment or as a multiple baseline experiment across several learning tasks or several learners. If the experimental procedures are ones which provide a reliable measure of experimental effects, this will demonstrate the reproducibility of these effects from the outset.

Until such time as an agreed experimental paradigm emerges in the scientific literature, direct replication should be considered to be a mandatory requirement of all experimental analyses of learning. Note that this requirement also solves the at present unresolved problem of what constitutes an adequate degree of experimental control in a learning experiment. Experimental procedures which, when replicated, produce closely similar effects, can be said to be reliable. The degree of experimental control which will need to be achieved if there is to be a science of learning is the degree of control which is found, through practical experience, to be necessary in order to produce reproducible measures of experimental effects.

In order to measure the reliability of an experimental effect, each element of the experiment must be reproducible. Note that reproducibility has two components. Firstly, there is the question of whether or not the procedures have been sufficiently well documented to allow another investigator to replicate the procedures used. Secondly, there is the question of whether or not replications of the original experiment actually yield the same result (Simkins, 1969).

The generality of the experimental results

Having identified a reliable (that is reproducible) relationship between a given instructional variable and a given change in the performance of certain learners, the next question to be asked concerns the generality of the observed effect. In earlier times the generality of an experimental effect used to be referred to as the "external validity" of the experiment (Campbell and Stanley, 1963). The generality of an experimental result refers to the extent to which a functional relationship which has been found to exist under one set of conditions may be expected to be observed under other conditions, or with other subjects.

Note that, until a given experimental effect has been shown to be a reliable, questions regarding the generality of that effect are irrelevant. Until a given instructional variable has been found to have a reproducible effect on learning under at least one set of conditions, it makes no sense to enquire about the range of conditions under which that effect may be expected to occur.

The generality of an experimental result depends upon the results of subsequent replication experiments which explore the limits of the generality of the original finding (Johnston and Pennypacker, 1980; Kazdin, 1980; Sidman, 1960). Replications which are designed to explore the generality of an observed relationship are usually referred to as "systematic replications" (Sidman, 1960).

If we want to know whether a given functional relationship applies across different types of learning tasks, or across learners with different learning histories, or across teachers, or across different types of instructional programmes, then the experiment must be repeated with different learning tasks, different learners, different teachers and different instructional programmes (Ballard, 1986; Barlow and Hersen, 1984; Borg, 1987; Gay, 1987; Greer, 1983; Johnston and Pennypacker, 1980; Kazdin, 1980; Sidman, 1960; White, 1984).

The result of a single experiment can not, in and of itself, have any

generality. Generality is something which arises only as a result of systematic replications of the original finding. This is as true for between-groups experiments as it is for within-subjects experiments. "Since statements of individual effect are usually not possible in the group-statistical design, such designs are generally incapable of assessing even the degree of generalized effects within the evaluation sample itself, let alone to some broader population" (White, 1984, p 117).

Discussions of the amount of work which is required to explore the generality of a given experimental result have been provided by Johnston and Pennypacker (1980) with respect to the effects of time-out from reinforcement, and by Barlow and Hersen (1984) with respect to the effects of differential social attention.

Scientific importance

The results of a series of experiments may or may not be of any scientific importance. The scientific importance of a functional relationship, once it has been identified, depends upon the degree of acceptance which it earns within the scientific community. A series of experiments may identify a functional relationship which can be reliably reproduced and which may even have some generality, but this finding may still be of little or no scientific importance. The findings of a series of experiments may be judged unimportant because the type of learning involved is trivial, or because the effect of the independent variable is a very weak one, or because the relationship is one of limited generality, or because the finding has no practical application. The cumulative development of a science provides the only final answer as to the scientific importance of particular experimental results (Sidman, 1960).

Part 3
Examples of the experimental analysis of learning

Attempts to design learning experiments which meet the technical requirements set out in the preceding section result in experiments which are very different to the between-groups experiments which have dominated the research on learning for most of this century. These differences are well illustrated by projects which have been undertaken by students enrolled in the author's Stage III course, EDUC 324: Individual Learning Processes. Since 1986, students enrolled in this course have been required to design and execute a simple learning experiment using any readily available subject. These student experiments are expected to meet three requirements.

First, the students must select, for their analysis, some kind of learning process. That is, they must analyse the acquisition of a behaviour or skill which their subject has yet to master and not a behaviour which the subject can already perform.

Secondly, students are expected to design a continuous measure of learning, that is, one which can be administered repeatedly as the experiment progresses. They are also expected to evaluate the reliability of the measurement procedure which they have devised.

Thirdly, the students are required to use a within-subject experimental design to measure the effects of a change in instructional procedure on rate of learning and to achieve a high degree of experimental control over their experimental treatments. Wherever possible, the students are also expected to measure the reliability of any treatment effects which they observe by including at least one replication within their experiment.

The design of these training experiments is guided by an instructional manual (Church and Liberty, 1992) which sets out the steps involved in the design of within-subjects analyses of learning. The students work through this manual during a series of 10, weekly, laboratory classes. A more detailed description of the teaching method, together with an evaluation of the course will be found in (Church, 1992a).

During the seven years that this course has been offered, students have completed about 200 simple learning experiments. It is not possible within the scope of this paper to review all of these studies. Instead, I have selected five pairs of studies to illustrate the kinds of procedures used, and the kinds of results which have been obtained.

The effects of modelling on rate of acquisition

Each year a number of students elect to study the effects, on rate of learning, of various showing and telling procedures. Figures 1 and 2 present the results of two such experiments (Richards, 1991; Powell, 1992). In both of these experiments, the student-researcher elected to measure the effects of introducing a teacher modelling condition during instruction in how to play a musical instrument.

In the Richards (1991) experiment, the subject was an 8-year old boy who was learning to play the recorder. The daily test consisted of an 8 to 10 bar tune containing all 10 of the notes being taught. The child's performance on the first 20 notes of the tune was assessed. There were three test tunes, the test tunes changed each day, and they were tunes which were never practised. During the modelling conditions, the teacher modelled correct fingering, modelled new notes, and modelled practice tunes during the practice phase of the lesson. During the no modelling conditions, the child had fingering diagrams to refer to but otherwise teaching was limited to verbal instructions. As can be seen from Figure 1, the rate of improvement on the test tunes was found to be greater during the modelling days than during the no modelling days.

In the Powell (1992) experiment, the learner was a 14-year girl who could read and play notes and note values from the treble clef and who wanted to learn to play the "left hand" notes so that she could play "proper tunes" on the piano. The daily test consisted of one of eight tunes. Each contained at least nine of the eleven left hand notes being taught and each contained at least 50 left hand notes to be played. The measure of improvement was the proportion of left hand notes played correctly and in tempo during playing of the test tune. The experimental treatments were similar to those used by Richards (1991). As can be seen from Figure 2, the rate of improvement in the playing of bass clef notes was greater during the modelling condition than it was during the days when the teacher was limited to verbal instruction.

The modelling effect has been replicated during cooking instruction with an eight year old girl (Thomas, 1991), dressage instruction with a 7-year old girl (Osborn, 1987), and instruction in machine oversewing and seam sewing with a 14-year old girl (Tipples, 1992).

The effects of prior verbal rehearsal on rate of acquisition

When faced with the task of teaching someone how to perform a new procedure, teachers commonly begin instruction with activities which are designed to teach the student to describe the steps which are involved in performing the task. This teaching procedure is based on the presumption that, once the student can remember (that is, describe) the steps which they need to perform, they will also have learned how to perform these steps.

Several student experiments have examined the effect, on rate of acquisition, of teaching students to describe the sequence of steps involved in completing a particular task before beginning to practice the task itself. The results from two of these experiments are presented in Figures 3 and 4. In both of these experiments, the subjects were taught to prepare and cook selected recipes. The measure of improvement was the proportion of steps in the preparation of the recipe which the student could perform without advice or assistance.

In the first experiment (Teo, 1991), a 13-year old boy was taught some of the basic skills which are involved in the preparation of Chinese dishes. In the Verbal Rehearsal Treatment, the student read, discussed, and described the steps to be taken in the preparation of that day's recipe before embarking upon the cooking of that recipe. In the Practice Only Treatment the student began straight away to perform each of the steps described in the recipe without any prior discussion. In both treatments, the teacher identified for the student any mistakes which were being made in the execution of each step and then described or showed the student how to perform the step correctly. The results, which are shown in Figure 3, reveal that the cooking tasks which were discussed before hand were performed more accurately than those which were not discussed and that this is true

throughout the experiment. The slope of the trends further indicate that there was no difference in the rate of improvement under either teaching condition.

In a closely similar experiment, Paltridge (1992) taught a 16-year old boy with mental retardation how to prepare and cook four simple recipes. For two of the recipes (Apple Surprise and Baked Potatoes) discussion of the steps to be taken continued until the student could describe these steps. Then the student began work on the recipe. The results of this experiment are shown in Figure 4. The trend lines for the four tasks are all parallel. This suggests that the recipes which were preceded by verbal rehearsal were mastered no more speedily than the recipes which were not preceded by any discussion. The tasks which were not discussed beforehand were of course performed much less well on the first day (as is to be expected) and this difference continued to exist throughout the experiment. But the rate of improvement on the tasks which were not discussed was no less that the rate of improvement on the tasks which were discussed beforehand.

The results of these two experiments confirm the widely held belief that verbal rehearsal confers a benefit during the learning of new procedures. Contrary to popular belief however, the results further suggest that this benefit occurs only during the very earliest training sessions. Verbal rehearsal did not result in a faster rate of mastery of the new procedures. Note also that the repeated measures design used in these experiments provided information about the rate of improvement - information which would have been lost if learning had been assessed only at the beginning and at the end of instruction.

Effects of rehearsal training on rate of acquisition of spelling words

The third group of experiments selected for discussion involved the evaluation of a a method for practising unknown spelling words - the "Spell-Write" procedure described by Croft (1983). The procedure is a self-rehearsal procedure.

Presented in Figure 5 are the results of an experiment by Lake (1989) in which two children were tutored in spelling at home. The first child, KL, was seven years of age and the second child, TL, was five years of age. Lake tutored these two children for 14 daily sessions - setting each child to learn their own set of eight new spelling words each day. The words were identified during pretesting as words which the child could not spell. The words which were practised each day were tested the following day. During the Before Training days, the children were simply instructed to practise the words. In KL's case this usually involved looking at the word, writing it down, and saying the letters while looking at the word. In TL's case this usually involved looking at the word and saying the letters out loud two or three times. Between Days 5 and 6, KL was taught to use the Spell-Write rehearsal procedure with each of the words to be learned. This was repeated for TL between Days 7 and 8. On subsequent days the children were checked to ensure that they were, in fact, using this procedure during practice sessions. The effect of the rehearsal training can be seen in Figure 5. The rate of acquisition of new spelling words increased from 3.4 to 7.4 words per session for KL and from 2.1 to 5.3 words per session for TL.

This experiment has now been replicated nine times (McBreen-Kerr, 1986; Rae, 1988; Ham, 1988; Longbottom, 1988; Lake, 1989; Tudor, 1992; Lake, 1992). The results of seven of these experiments will be found in Church (1990). In each of these experiments (except the second of the Longbottom experiments and the experiment by Lake (1992)) the rate of acquisition following rehearsal training increased at least two-fold. These replications involve pupils of varying ages from 5 years through to 13 years of age. They include pupils whose school achievement is well above average (Lake, 1989) and pupils whose school achievement is well below average (Longbottom, 1988; McBreen-Kerr, 1986). The effect has been obtained by seven separate investigators of whom only three were teachers.

These experiments are instructional in a number of respects. They demonstrate the measurement of learning using an equal-interval scale with a true zero - the kind of scale which is necessary if the results of several experiments are to be compared. They demonstrate that this particular instructional intervention results in at least a two-fold increase in rate of learning. They demonstrate that it is possible to design learning experiments which have reproducible effects even when these are carried out under relatively naturalistic field settings. And they demonstrate, much more economically than would have been possible using a between-groups design, that this particular finding is likely to be one with very considerable generality.

Effects of error feedback on rate of acquisition

A number of students have chosen to study the effects of feedback on rate of acquisition. In one of the first of these (McIlhone, 1987), a teacher provided individualized daily tuition in printing skills to a 6-year old boy. The daily lessons consisted of practice on two to four new letters each day. Changes in printing skill were recorded from day to day using acetate overlays to assess the accuracy with which the letters in a 6- to 8- word test sentence had been copied. The procedure was similar to that described by Trap, Milner-Davis, Joseph and Cooper (1978). The same set of eight test sentences were used during both the first and second experimental treatments to control for task difficulty. During the first experimental treatment, letters (in the test sentence) which had been printed satisfactorily were ticked and the child praised for his efforts. During the second experimental treatment, the acetate overlays were used in the presence of the child, letter forms which were not yet satisfactory were identified for the child, and the teacher explained why the unsatisfactory letters had not been ticked. The effect of this additional feedback on rate of acquisition Is shown in Figure 6.

In another experiment in this group, Yee (1992) examined the effects of error feedback during keyboarding tuition using a typing tutor with a 15-year old girl. The typing tutor was one which normally provided a "beep" if the student typed an incorrect character during the course of practice exercises. However, this facility was one which could be turned off by the teacher if desired. Figure 7 shows the rate of improvement in typing accuracy which was observed during blocks of

practice sessions in which the audio error feedback was (a) turned on and (b) turned off. The figure shows both the number of correct and the number of incorrect typing responses during 1 minute tests. As can be seen from the figure, the student's typing improved more rapidly during those session blocks in which audio error feedback was provided.

A marked decline in rate of progress was also observed by Murphy (1986) when error feedback was withdrawn following daily typing tests with a 12-year old girl and by Watkins (1989) when information about corrects and errors was withdrawn during daily tests of the ability to read and play 20 notes on a piano keyboard. The Murphy experiment is described in Church (1992b) and the Watkins experiment in Church (1990).

The results of these four experiments suggest that the provision of error feedback during practice results in much faster rates of acquisition than is the case when this information is absent. The results of these studies also indicate that the effect of this kind of feedback is likely to be a fairly robust effect and that it is likely to be observed across a variety of different kinds of learning tasks.

Effects of extrinsic reinforcement on rate of learning

EDUC 324 students frequently elect to study the effects of various kinds of rewards on acquisition. The results of two of these experiments will be found in Figure 8 (Yee, 1989) and Figure 9 (Farrow, 1991).

In the Yee (1988) experiment, a 4-year old boy was tutored in letter recognition (learning to say the names of the lower and upper case letters of the alphabet). Five letters were practised each day. The letters which had been practised one day were tested the following day. Letters correctly identified during each test were removed from the practice set and replaced by new letters. The letters were divided into two sets. On even numbered days, the child practiced letters from Set A and each correct practice response earned the child a wine gum. The results, which are shown in Figure 8, show that the child learned the Set A letters more quickly than the Set B letters.

A closely similar result was obtained by Calnan (1988) using letter recognition as the learning task, jellybeans for correct responses as the reward, and a multiple baseline across two 4-year olds as the experimental procedure.

In the Farrow (1991) experiment a 3-year old girl was tutored in letter recognition using games of various kinds. During the playground reinforcement phases of the experiment, improvements in the child's performance on the daily test resulted in the child being taken to a nearby adventure playground to play. The child's ability to name letters was assessed by spreading out the 26 letters in front of the child and counting the number of letters which the child could name in 90 seconds. As can be seen from Figure 9, the child made a faster rate of progress during those phases when the playground reinforcement procedure was operating.

Attempts to replicate these experiments with older children have failed in almost every case. Regardless of the form of the reward which is offered, introduction of the opportunity to earn a reward has not resulted in any detectable increase in rate of acquisition in any of the studies involving older children (e.g. Whiteford, 1986; Harris, 1987; Horne, 1989; Kincaid, 1989; Nairn, 1990; Roberts, 1990; Forrest, 1991). It seems that with older children, additional extrinsic rewards do not have an effect over and above the reinforcement which arises naturally as a result of improvements in skill. Note, however, that these experiments do not report an undermining effect either. The introduction of additional rewards over and above the reinforcement which results from improvement appears neither to accelerate nor to decelerate rate of acquisition. It simply has no effect.

Part 4
Discussion

The experiments described in the preceding section show that it is possible to design learning experiments which satisfy the technical requirements described in Section 2. In each experiment, the student succeeded in designing a demonstrably reliable measure of rate of learning, the measures proved to be sensitive to experimental effects, the measurement scales were equal-interval scales with a true zero, and the experimental effects which were observed proved to be reproducible.

These experiments differ in a number of respects from the between-groups experiments which continue to dominate the literature on learning. These experiments do not employ samples of subjects, they do not limit the assessment of learning to a single observation, they do not measure experimental effects against the within-groups variance, and they do not use probability statistics to evaluate the importance of treatment effects. Experiments such as these demonstrate that there are experimental procedures other than the between-groups procedure which can be used to analyse learning. In demonstrating an alternative to the between-groups procedure, experiments such as these raise a number of critically important questions regarding the kind of experimental procedure which is most likely to advance our understanding of the learning process.

The first question which learning researchers must ask of their experiments is this. Are we trying to understand the behaviour of groups or are we trying to understand the behaviour of individuals?

The experiments reported in the preceding section demonstrate that it is possible to design reliable measures of the effects of particular instructional variables on the rate of learning of individuals. However, the majority of learning experiments continue to be designed as between-group experiments. Johnston and Pennypacker (1980) argue that if there is to be a science of learning, then its aim must be to explain and to predict the learning of individuals. The performance of an individual and the performance of a group are two fundamentally different phenomena and the investigator must decide which of these phenomena he/she is interested in studying. "In every area of science there are critical times when such decisions must be made, and the consequences are far-reaching. If the correct decision is generally accepted, the science will advance. If the incorrect choice is adopted, the science will experience a period of stagnation until the situation is righted" (Sidman, 1960, p 54).

Secondly, we need to think very carefully about the following question. Should inter-subject differences be conceptualized as error variance or are they in fact the very phenomena which we are trying to understand and to explain?

All teachers know that teaching has to be adapted to meet the needs of the individual pupil and that a particular teaching procedure may have different effects with different learners. The reinforcement experiments cited in Section 3 remind us that the effects of an instructional procedure may vary with the age of the learner. Between-groups experiments treat individual differences in performance following instruction as "error variance", or "background noise" (Simkins, 1969). This background noise is then attributed to "chance". However, this "chance" can only be due to the combined effects of uncontrolled variables. "If such variables are in fact controllable, then chance in this sense is simply an excuse for sloppy experimentation. If the uncontrolled variables are actually unknown, then chance is, as Boring has pointed out, a synonym for ignorance" (Sidman, 1960, p 45). Johnston and Pennypacker (1980) argue that the goal of a science of learning is to explain, not to passively accept, individual differences. "The subject matter of all science is variability. ... Without variability there would be no science, as there would be no need for classification and organization and nothing to explain. ... Like all other natural phenomena, behavior displays variability, and, like the other sciences, the science of behavior has as its task the explanation of that variability. For a variety of reasons, recognition of this mission or even acceptance of its possibility has been slow in coming" (Johnston and Pennypacker, 1980, p 202).

The third question to be asked concerns the most appropriate way of operationalizing and measuring learning.

The experiments described in the preceding section each selected a readily identifiable human skill as an exemplar of learning and operationalized these in a way which produced a measure of the speed or rate of improvement in the performance of those skills. In contrast, the majority of between-groups experiments operationalize learning using tests of recall, retention, or achievement rather than rate of learning. If we are measuring learning in individuals, then it seems appropriate to measure instructional effects in terms of the rate of performance change which they produce. However, it needs to be recognized that there is no a priori analysis which can be used to determine how learning should be operationalized in the sense that its controlling variables can be discovered by experimental analysis. What constitutes an appropriate unit of behaviour change for experimental analysis is an entirely pragmatic matter. If the behavioural dimension selected for analysis is one which proves to be reliably related to certain independent variables then its continued use as a dependent variable is justified. If the behavioural dimension selected for analysis cannot be shown to be reliably related to any independent variables then its continued use as a dependent variable is pointless (Johnston and Pennypacker, 1980). "If we cannot establish any functional relationships between independent variables and the dependent variable then we are forced to the conclusion that we must be using the wrong measurement operation" (Simkins, 1969, p 132).

Fourthly, we need to consider the characteristics of useful measurement units. Should the changes which we refer to as learning be quantified using relative or absolute units of measurement?

In each of the experiments reported in the preceding section, rate of learning has been measured using an absolute scale of measurement. In the rehearsal training experiments, for example, the investigator counted the words which the child could (a) not initially spell and which they could (b) still spell correctly 24 hours after they had been practised. This measurement operation produces an equal interval scale with a true zero. A rate of 8 words correct per day represents a learning rate which is twice as great as a rate of 4 words correct per day. The measurement operation yields a measurement result which can be operated on arithmetically and which can be directly compared with results obtained in other experiments using the same measurement operation. The measurement units most commonly used in between-groups experiments are (a) the "statistically significant effect" (which has only two values) and (b) the "effect size". Both of these measurement units are relative, not absolute units of measurement. A given difference between two group means can yield a range of effect sizes depending upon the degree of variability in the scores of individual subjects. The history of science tells us that very little progress is made until those active in the field succeed in developing absolute and standard units of measurement, such as the ratio scales used to measure length, time, weight and density. "The definition of a standard unit of measurement is an important determinant of the reliability and reproducibility of measurement operations" (Simkins, 1969, p 134). Finally, we need to ask whether any conclusions at all should be drawn from the results of a single unreplicated experiment of unknown reliability and generality.

Generally speaking, a within-subjects experiment will not be accepted for publication until its results have been replicated. The journals which publish between-groups studies, on the other hand, rarely require a replication as a condition of publication. This state of affairs is justified by the belief that a "significant treatment effect" is a reliable effect. I believe that the major reason why the results of between-groups analyses of learning have failed to identify generally agreed "principles of learning" is because those who have worked within this paradigm have ignored the need for both the direct replication and the systematic replication of their experiments. Had this been a requirement of publication, then those working within this paradigm would long ago have discovered which of their measurement procedures were reliable, which of their experimental procedures were reliable, and which of their experimental results were reliable.

One of the distinguishing characteristics of a science is a cumulative development in both the investigative methods which are being used and in the experimental results which are being reported. The detailed critiques which have been provided by such writers as Sidman (1960), Simkins (1969), Johnston and Pennypacker (1980), Greer (1984) and others lead fairly directly to the conclusion that the between-groups paradigm which has dominated learning research for three-quarters of a century is essentially a pre-scientific endeavour which will never identify the variables of which learning is a function. The shortcomings of the between-groups approach to experimentation include the use of changes in the mean performance of groups as the primary datum (rather than changes in the performance of individual learners), the use of only one or two observations of performance change (instead of regular observations of performance change), the use of measures of achievement which mix items measuring improvement with items which could not possibly measure improvement, the failure to develop standard measurement units which would allow the results of different experiments to be directly compared, the failure to demonstrate the reliability of the procedures used to measure behaviour change, and the failure to demonstrate the reliability of treatment effects. (This is not to say that group designs have no place in educational research. These designs are, of course, appropriate where the question of interest concerns the relative achievement levels of groups or populations of students who have been exposed to different educational provisions. But questions regarding the conditions which affect the skill development of individual learners cannot be answered using these designs.)

One of the advantages of a paradigm shift to a within-subjects analysis of learning processes is that such experiments are inherently more economical of investigator effort than is the case with the between-groups experiment. This is because the experimenter is in direct contact with the phenomenon of interest (changes in the learner's behaviour) and can make adjustments to his or her experiment as the experiment proceeds.

A second advantage of the within-subjects methodology is that it takes much less time to train a competent experimenter than is the case with the between-groups methodology and the complex statistical analyses which such "experiments" entail. Less than a year's training seems to be sufficient to bring the majority of graduate students to the point where they can design a reasonably adequate within-subjects experiment and even undergraduate students have shown that they can master the methodology. Acquiring mastery of the design principles and the statistical analysis procedures required for the design and execution of a between-subjects experiment takes very much more training time and sometimes proves to be beyond the ability of even graduate students.

In every country, the number of persons involved in the study of learning processes is but a tiny fraction of the number of persons employed as teachers and the money applied to educational research is but a tiny fraction of the money applied to schooling (Church, 1989). Unless the research base can be broadened, improvements in our understanding of teaching and learning and the relationships between the two is likely to be painstakingly slow. A third advantage of the within-subjects methodology, as I have stated elsewhere (Church, 1975; 1992a), is that it brings the techniques for measuring the effects of teaching strategies on pupil learning within the grasp of those who have the most to gain by their application - teachers themselves.

The experiments reported in this paper, and others like them, demonstrate that a within-subject methodology can be used to identify observable conditions which affect the rate of acquisition of new behaviours, skills, and competencies. It remains to be seen, of course, whether or not the controlled within-subject analysis of learning processes will succeed in identifying functional relationships with any more reliability than has occurred with the application of the between-groups paradigm.

References

Ballard, K.D. (1986). Group designs, within-subject designs, case-study designs, and qualitative methodologies in educational and psychological research. New Zealand Journal of Educational Studies, 21, 42-54.

Barlow, D.H. and Hersen, M. (1984). Single-case experimental designs: Strategies for studying behavior change (2nd ed.). New York, Pergamon Press.

Borg, W.R. (1987). Applying educational research: A practical guide for teachers. New York: Longman.

Calnan, B. (1988). The use of a reinforcer system to facilitate children's acquisition of the alphabet. Unpublished Educ 323 Field Study report, University of Canterbury, Education Department.
Campbell, D.T. and Stanley, J.C. (1963). Experimental and quasi-experimental designs for research. Chicago: Rand McNally & Co.
Church, R.J. (1975). Could teachers be doing worthwhile research? set, No 1, Item 5.
Church, R.J. (1976). A study of the components of an effective teaching strategy. Unpublished PhD thesis, University of Canterbury.
Church, R.J. (1989). Educational research: The state of the science. Input, 11(1), 1-3.
Church, R.J. (1990). The use of within-subjects designs to measure the effects of teaching on learning. Paper presented to the 1990 conference of the N.Z. Association for Research in Education, Auckland, December, 1990.
Church, R.J. (1992a). Teaching behaviour analysis research skills to undergraduates. Paper presented to the Fourth World Congress on Behaviour Therapy, Gold Coast, Australia, July, 1992.
Church, R.J. (1992b). Measuring learning. Paper presented to the Second Joint AARE/NZARE Conference, Geelong, Australia, November, 1992.
Church, R.J. and Liberty, K.A. (1992). The experimental analysis of learning. University of Canterbury, Education Department.
Cooper, J.O., Heron, T.E. and Heward, W.L. (1987). Applied behavior analysis. Columbus: Merrill Publishing.
Croft, C.O. (1983). Teachers Manual for Spell-Write. Wellington, New Zealand Council for Educational Research.
Farrow, G. (1991). The effects of reinforcement on letter sound acquisition. Unpublished Educ 324 Field Study report, University of Canterbury, Education Department.
Forrest, K. (1990). The effects of reinforcement in the form of a 15 minute game on a child's progress when learning to type on a word processor. Unpublished Educ 323 Field Study report, University of Canterbury, Education Department.
Gay, L.R. (1987). Educational research: Competencies for analysis and application. (3rd ed.). Columbus: Merrill Publishing Company
Greer, R.D. (1983). Contingencies of the science and technology of teaching and pre-behavioristic research practices in education. Educational Researcher, 12, 3-9.
Ham, J.M. (1988). The effects of the introduction of a new method for learning spelling on spelling achievement. Unpublished Educ 323 Field Study report, University of Canterbury, Education Department.
Haring, N. and Eaton, M. (1978). Systematic instructional procedures: An instructional hierarchy. In N. Haring, T. Lovitt, M. Eaton and C. Hansen (Eds.) The fourth R: Research in the classroom. Columbus, Charles E. Merrill.
Harris, A. (1987). The effects of a progress chart and Charlie Brown poster on the rate of learning of new spelling words. Unpublished Educ 323 Field Study report, University of Canterbury, Education Department.

Horne, C.A. (1989). The effect of edible reinforcement on a child's progress when learning to play the recorder. Unpublished Educ 323 Field Study report, University of Canterbury, Education Department.
Kazdin, A.E. (1980). Research design in clinical psychology. New York: Harper and Row.
Kincaid. A. (1989). The effects of verbal and edible rewards on learning how to play the recorder. Unpublished Educ 323 Field Study report, University of Canterbury, Education Department.
Johnston, J.M. & Pennypacker, H.S. (1980). Strategies and tactics of human behavioral research. Hillsdale, NJ: Erlbaum.
Lake, C.M. (1989). The effect of an instructional package to enhance spelling performance in children. Unpublished Educ 323 Field Study report, University of Canterbury, Education Department.
Lake, Calvin. (1992). The effects of the Spell-Write rehearsal strategy for learning words on the spelling accuracy of an 8-year old girl. Unpublished Educ 324 Field Study report, University of Canterbury, Education Department.
Longbottom, G.C. (1988). The effect of teaching a structured method of study on spelling acquisition. Unpublished Educ 323 Field Study report, University of Canterbury, Education Department.
Martin, T. and Church, J. (1992). The Co-operative Reading Resource and how it changed the reading skill of 8- and 9-year olds. set, No 2, Item 6.
McBreen-Kerr, S. (1986). Introducing a specific learning strategy to the learning of spelling words in a remedial pupil. Unpublished Educ 323 Field Study report, University of Canterbury, Education Department.
McIlhone, H. (1987). The effects of feedback on basic script letter formation. Unpublished Educ 323 Field Study report, University of Canterbury, Education Department.
Murphy, J. (1986). The effects of feedback and non-feedback on the teaching of touch typing. Unpublished Educ 323 Field Study report, University of Canterbury, Education Department.
Nairn, D. (1990). The effects of praise with reinforcement and praise without reinforcement on learning to tell the time. Unpublished Educ 323 Field Study report, University of Canterbury, Education Department.
Osborn, V. (1987). The effects of differing degrees of guidance on a learner's acquisition of a dressage test. Unpublished Educ 323 Field Study report, University of Canterbury, Education Department.
Paltridge, J. (1992). The effects of prior and during instruction on the completion of four recipes. Unpublished Educ 324 Field Study report, University of Canterbury, Education Department.
Powell, M. (1992). The effects of verbal instruction and physical demonstrations on the ability to learn tunes on the piano. Unpublished Educ 324 Field Study report, University of Canterbury, Education Department.
Rae, A. (1988). The effects of the 'Spell-Write' rehearsal strategy on increasing spelling performance. Unpublished Educ 323 Field Study report, University of Canterbury, Education Department.
Richards, L. (1991). The effects of teacher modelling and diagrams on the

acquisition rate of notes on a recorder for an 8-year old with no previous knowledge of the instrument. Unpublished Educ 324 Field Study report, University of Canterbury, Education Department.
Roberts, N. (1990). The effect of verbal reinforcement and a star chart on the acquisition of fact propositions. Unpublished Educ 323 Field Study report, University of Canterbury, Education Department.
Sidman, M. (1960). Tactics of scientific research. New York: Basic Books.
Simkins, L.D. (1969). The basis of psychology as a behavioral science. Waltham, Blaisdell.
Teo, Wee Keng. (1991). The effects of verbal rehearsal on the rate of acquisition of new motor response chains by a 13-year old boy. Unpublished Educ 324 Field Study report, University of Canterbury, Education Department.
Thomas, A. (1991). The effects of demonstrations on the acquisition of cooking skills. Unpublished Educ 324 Field Study report, University of Canterbury, Education Department.

Tipples, R. (1992). The effects of oral description only and oral description plus demonstrations on the accuracy and fluency of sewing a garment using four specific stitches on a sewing machine. Unpublished Educ 324 Field Study report, University of Canterbury, Education Department.
Trap, J.J., Milner-Davis, P., Joseph, S. and Cooper, J.O. (1978). The effects of feedback and consequences on transitional cursive letter formation. Journal of Applied Behavior Analysis, 11, 381-393.
Tudor, R. (1992). The effects of the Spell-Write method of learning a word on spelling achievement. Unpublished Educ 324 Field Study report, University of Canterbury, Education Department.
Watkins, A. (1989). The effects of a progress chart on the time taken to learn to play 10 notes on the piano keyboard from written music. Unpublished Educ 323 Field Study report, University of Canterbury, Education Department.
West, R.P., Young, K.R. and Spooner, F. (1990). Precision teaching: An introduction. Teaching Exceptional Children, 22(3), 4-9.
White, O.R. (1984). Selected issues in program evaluation: Arguments for the individual. Advances in Special Education, Volume 4, pp 69-121.
White, O.R. and Liberty, K.A. (1976). Behavioral assessment and precise educational measurement. In N.G. Haring and R.L. Schiefelbusch (Eds.), Teaching special children. New York, McGraw-Hill Book Company.
Whiteford, T. (1986). The effect of reinforcement on a child's progress when learning to type. Unpublished Educ 323 Field Study report, University of Canterbury, Education Department.
Yee, I. (1989). The effects of feedback and reinforcement on rate of acquisition of the names of letters by a 4-year old. Unpublished Educ 323 Field Study report, University of Canterbury, Education Department.
Yee, B. (1992). The effects of immediate digitized audio feedback on the acquisition of touch typing. Unpublished Educ 324 Field Study report, University of Canterbury, Education Department.