# MEASURING LEARNING

John Church
Education Department
University of Canterbury

This paper identifies the kinds of behaviour changes which are commonly included under the heading of "learning". It argues that the term refers to the behaviour changes of individuals and that behaviour can change along any of several dimensions. The paper also identifies the main technical requirements which need to be met if learning is to be measured in a scientific manner. These include measurement accuracy, measurement sensitivity and the use of standard units to report measurement results. When learning is tested only once or twice, these requirements are almost never met. The case for continuous measures of learning is illustrated using a variety of student-administered learning experiments.

The investigator who wishes to identify variables which affect student learning must first develop a measure of learning. In the great majority of learning experiments conducted this century, learning has been measured by developing a test of learning outcomes which is then administered either once (at the end of instruction) or twice (at the beginning and at the end of instruction). This paper agues that this procedure may be adequate to the task of measuring achievement but that it is an extremely unsatisfactory way of measuring learning.

## Part 1
## Characteristics of the phenomenon to be measured

Learning is said to have occurred when a learner demonstrates that they can do something which they could not do before, or when a learner demonstrates that they can now do something better than they could before. Before we can develop accurate measures of learning we must be clear about the nature of the phenomenon which we are attempting to observe and record.

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

### Learning is an individual phenomenon

The first observation which we can make is that it is individuals who learn. If two pupils are exposed to the same instruction and one becomes more skilled or knowledgeable and the other does not, we say that the first pupil has learned and that the second pupil has not.

Because learning is an individual phenomenon it follows that the measures of learning which are used in learning experiments must be measures which provide an accurate record of the learning of individuals. If an experiment involves more than one individual, then the measure of learning must provide an accurate record of the learning of each individual. A mean

score can never function as a measure of learning because a mean score is never an accurate representation of the learning of each of the individuals who took part in the experiment. In fact a mean score may not accurately represent the learning which occurred with respect to any of the individuals in the group.

Learning involves change

When an individual learns to do something which they could not previously do, their performance changes - sometimes permanently. Johnston and Pennypacker liken this change to the motion of matter in space: "Behaviour possesses many of the characteristics of matter in motion and the same principles of measurement are applicable" (Johnston and Pennypacker, 1980, p 73). West, Young and Spooner liken this change to a motion picture: "Learning is a continuous process; it can be compared to a motion picture that gives a continuous study of an event or series of events. ... If we can study many frames from the motion picture, we will have a much better idea of the continuous process the motion picture represents" (West, Young and Spooner, 1990, p 5-6).

This observation has profound implications for the measurement of learning - implications which are only gradually coming to be recognized.

First, one cannot measure change by making only a single observation. An experiment which simply measures pupil achievement at the end of an experimental sequence of instruction is not a learning experiment because it can provide no information regarding the way in which the learner's behaviour has changed.

Secondly, an experiment which observes student performance on only two occasions - such as before and after a sequence of instruction - provides the weakest possible measure of learning because it collects the smallest possible amount of information regarding performance change. Any measurement strategy which measures performance more than twice will provide a more detailed picture of learning.

Thirdly, frequent and repeated observations of performance provide the most detailed picture of the changes which we refer to as learning. "Learning requires more than a single instance of measurement. In fact, the more instances of measurement we have to inspect, the more accurate and representative will be our interpretation of learning. If we can study many frames from the motion picture, we will have a much better idea of the continuous process the motion picture represents" (West, Young and Spooner, 1990, p6). Following the motion picture analogy, attempts to measure performance change by means of the repeated application of a given measure of performance will be referred to as continuous measures of learning to distinguish them from "one shot" measures of achievement and attainment.

Fourthly, in order to observe performance repeatedly, the measurement procedure must be one which can be applied repeatedly during the acquisition period. This has a number of practical implications which will be discussed below.

Only changes in performance can be measured at the present time

The conclusion that learning is occurring can only be based upon the

observation that the learner's performance is changing - most commonly in the direction of increasingly competent performance of some skill, or increasingly competent completion of some task, or increasingly competent performance on some kind of test or examination.

It is recognized that many people respond to the observation of an improvement in performance with the inference that the learner has now acquired some new "ability", "knowledge", "understanding", "cognitive structure", or "metacognitive skill" (to use some common metaphors). I don't think that there is anything to be gained by debating whether learning involves a change in performance (which it clearly does) or whether learning involves a change in the brain of the student (which is probably also the case). It is sufficient to note at this stage in the development of psychology as a science that changes in performance can be accurately measured using current measurement techniques whereas changes in the brain cannot. Procedures for measuring brain changes during learning have yet to be developed. If people wish to continue using metaphors like "cognitive structure" little harm will be done provided we do not lose sight of the fact that changes in the learner's performance can be observed and recorded whereas, at the present time, changes in the learner's state of mind cannot.

My own view is that it is possible to observe, measure and analyse changes in performance without making mentalistic inferences. I would further argue (probably without convincing anyone) that it is more parsimonious (and hence more scientific) to do so.

Performance can change along any of several dimensions

Not only does learning lead to a change in performance, but this change may occur with respect to any one of a number of different aspects or dimensions of performance (Haring and Eaton, 1978; Johnston and Pennypacker, 1980; West, Young and Spooner, 1990; White and Liberty, 1976).

For example, the ability to perform completely new responses may be acquired - as when there is an increase in the number of items of clothing which the learner can put on unaided, or an increase in the number of letters which the learner can print unaided. This dimension of behaviour change is often referred to as "skill acquisition" or "knowledge acquisition".

Secondly, the ability to respond correctly may be acquired. For example, the number of multiplication facts which the learner can complete correctly, or the number of test questions on a given topic which the learner can answer correctly may increase. This dimension of behaviour change is often referred to as an improvement in "knowledge" but I prefer the less mentalistic term "accuracy".

Thirdly, the ability to correctly apply a particular, previously learned, response or skill to a new situation may be acquired. For example, a child who can apply a rule to solve problems which are set out in a particular way may learn to apply the same rule to solve new problems which are set out in a completely different way. This dimension of behaviour change used to be referred to as "transfer" but is now more commonly referred to as

"generalization".
Fourthly, the ability to respond with greater speed or automaticity may be acquired. For example, the speed with which words can be read, or written, or spelt correctly, may increase. This dimension of behaviour change used to be referred to as "overlearning" but is now more often referred to as "fluency".
Fifthly, the ability to respond for longer periods of time may be acquired. For example, the length of time for which the learner can maintain a given reading speed, or writing speed, or swimming speed may increase. This dimension of behaviour change may be referred to as "endurance".
These distinctions are important because earlier and later phases of instruction in any given skill tend to target different dimensions of behaviour change. Haring and Eaton (1978), for example, observe that, for any definable skill, the student must first acquire the skill (i.e. learn how to perform the behaviour and when to perform it), then become fluent in performing the skill (i.e. learn to perform the behaviour at an appropriate speed), and finally learn to generalize and to adapt the skill to new situations.
These distinctions are also important because the kind of measure which is appropriate depends upon the behavioural dimension which is changing (Johnston and Pennypacker, 1980).
When someone acquires the ability to perform a new behaviour or skill, the topography of their behaviour changes, that is, they acquire the ability to move in a new way. A variety of procedures have been proposed for measuring changes in this dimension of behaviour. One commonly used procedure is to ask raters to grade the sophistication of the learner's performance against specified criteria. This procedure suffers from a very serious weakness. To provide a detailed picture of change we need a scale with many categories on it (20 or 30 rather than 3 or 5). However, as the number of categories increases, inter-rater reliability decreases dramatically.
A much more accurate way of quantifying the emergence of a new behaviour is to count the number of responses which meet a certain standard in terms of their effect on the environment (White and Liberty, 1976; White and Haring, 1980). "We may describe behaviour in terms of critical effect, that is, the most important change which the behavior produces in the environment of the child. In walking, the most important change (critical effect) is moving the child from one place to another. Improvement in walking may be measured in terms of the number of times the child is able to achieve that critical effect. How many times can the child walk from point A to point B, in say, 10 minutes?" (White and Liberty, 1976, p 36). This procedure can be applied even to very complex performances (e.g. the number of reports written which are of publishable quality). Increases (from one observation to the next) in the number of responses which produce a given critical effect (or which can be performed to a given standard) provide a measure of improvement in this dimension of performance.
Accuracy (the ability to respond correctly) can be measured by counting the number of responses to a given number of test stimuli which meet the criteria for "correct responding". Increases (from one observation to the next) in the number of correct responses provide a measure of improvement

in this dimension of performance. Several writers (e.g. West, Young and Spooner, 1990; White and Haring, 1980) have made the point that the

measurement of accuracy requires the observer to keep a separate record of (a) correct responses, (b) incorrect responses (errors) and (c) skips (that is, items which are not attempted by the learner). This is because these three classes of behaviour are likely to be under the control of different kinds of events. Errors, for example, may be a function of the fact that the student is applying the wrong rule whereas skips may be a function of the fact that the student has not yet learned the rule which describes how to respond to these items.

Generalization (the ability to apply a given skill to new situations) is also measured by counting the number of correct, incorrect and skipped responses to test stimuli. The only difference is that the measurement of generalization requires test stimuli which have not been used during instruction and which have not been seen before by the learner.

Fluency can be measured by counting the number of correct responses provided by the learner during a given period of time (such as one minute). It can also be measured by recording the time which the learner takes to complete a procedure or task, or to produce a specified number of correct responses. Increases (from one observation to the next) in the number of correct responses which can be performed per unit of time provide a measure of improvements in fluency.

Fluency is a teaching goal with respect to all of the basic academic skills which are taught in schools (Lentz, 1988; West, Young and Spooner, 1990; White and Haring, 1980). While fluency is not always measured in the classroom, it is almost always appropriate to measure this dimension of behaviour during a learning experiment. This is because, unlike accuracy (which tends to be acquired fairly quickly), fluency is a dimension of performance which can continue to improve over quite long periods of time (e.g. Crossman, 1959). A measure of fluency provides a much more sensitive measure of learning than does a measure of accuracy (Howell and Lorson-Howell, 1990).

Changes in endurance can be measured by recording the period of time over which the learner is able to sustain a given level or rate of performance. Increases (from one observation to the next) in the length of time over which a given rate can be sustained provide a measure of improvement in endurance.

Pennypacker argues that measures of accuracy, fluency and rate of change are independent dimensions of performance. "We have repeatedly established that under these conditions, the correlation among these three measures are essentially zero, meaning that frequency, accuracy and accuracy change are independent of each other and are probably measuring different aspects of the behavioural process that occur during instruction. ... Because these measures are independent, we can combine them in various ways to generate composite measures of productivity that allow for all possible patterns of student performance change" (Pennypacker, 1976, p314).

Pennypacker also argues that, until more is known about the effects of instructional events on the various dimensions of performance, all

dimensions should be observed and recorded so that no information of potential value is overlooked (Johnston and Pennypacker, 1980).

## Part 2
## The technical requirements of good measurement

Measurement procedures which are developed for the scientific study 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 phenomenon of interest, which (b) are reproducible, which (c) produce results which can be reported in standard units, and which (d) are sensitive enought to detect changes in the dimensions of interest (Johnston and Pennypacker, 1980; Simkins, 1969).

### Accuracy of the measurement result

The paramount consideration during any attempt to measure learning in a scientific manner is that the measurement procedure should provide an accurate representation of the behaviour changes which occurred (Cone, 1981; Johnston and Pennypacker, 1980; Lentz, 1988). "The goal of scientific measurement is to arrive at the best possible estimate of the true value of some dimensional quantity. To the extent that this goal is achieved, the measure is said to be accurate" (Johnston and Pennypacker, 1980, p190). If a child is reading at a rate of 60 words per minute, and the measurement procedure records this as a rate of 60 words per minute, the procedure has produced an accurate result.

In order to demonstrate that a given measure of behaviour is accurate, some independent source of knowledge about the true state of affairs (e.g. the true behavioural frequency) must be accessible. The measurement result must be independently verifiable (Cone, 1981). This can be achieved in a variety of ways. For example, a videotape might be scripted to provide an exact and known number of instances of a given behaviour (such as a reading error rate of 1 in 20). If the measurement procedure involves the use of live observers making running records of reading, the accuracy of the results generated by this procedure can be readily checked by applying the procedure to the taped behavioural samples with the known error rates.

In order to obtain an accurate measure of some behavioural phenomenon it is first necessary to develop a reliable measurement procedure. A measurement procedure is a reliable procedure if, each time it is applied to the same phenomenon or state of nature it generates the same measurement result (Johnston and Pennypacker, 1980; Simkins, 1969). For example, if we wanted to obtain a set of accurate measures of the frequency with which a child is self-correcting during oral reading, the observer would need to be able to reliably (i.e. consistently) classify instances of correctly read words, incorrectly read words, and self-corrections.

Note, however, that the converse is not true. Demonstrating that the measurement procedure is reliable, does not demonstrate that the measurement result is accurate. It is possible for a measurement procedure

to generate a consistent result, but for that result to be consistently inaccurate (Cone, 1981; Johnston and Pennypacker, 1980). For example, a reading teacher who consistently misclassified re-runs as reading errors would produce measures of reading error rate which were consistently (i.e. reliably) inaccurate.

When a new measurement procedure is developed, it is important that the question "Will this procedure produce accurate results?" be answered prior to its use in any experiment. There is little point in investing time and money in a learning experiment until it has been demonstrated that the proposed measures of learning are ones which will produce an accurate result.

Reproducibility

Scientific data are data which are independently verifiable. To be scientifically useful, a measurement procedure must also be reproducible. "Any measurement procedure worthy of the name must have ... a well-specified operation or rule which can be communicated and applied by any sufficiently trained person" (Hays, 1967, p5). In order for a measurement operation to be reproducible, it must be described sufficiently explicitly for it to be reproduced by other investigators. This involves specification of the conditions under which measurements are to be made, specification of the procedure for making the measurements, and specification of the quantities which are being used to report the measurement results (Simkins, 1969).

Comparability

If a given phenomenon is going to be measured on more than one occasion or by more than one experimenter, and the results communicated to other investigators, or compared in some way, then the measurement results must be reported in standard units. Otherwise the measurement results have no meaning to anyone apart from the original investigator. "Scientific measurement of natural phenomena is a process involving quantification of observations with respect to a reference scale composed of and defined by units that are both absolute and standard" (Johnston and Pennypacker, 1980, p55). "Absolute units are those whose definition is independent of the measurement operation - although not necessarily of the measuring device" (Johnston and Pennypacker, 1980, p120). Standard units are necessary for both comparison and prediction. It is not possible to compare the results of two or more measures of the same phenomenon unless they are given in the same units.

The kind of unit which is most appropriate depends upon the behavioural dimension which is to be quantified. Skill level can be quantified by counting the number of demonstration responses which do and do not have a certain critical effect or which do and do not meet a specified standard. Accuracy level can be quantified by counting the number of responses which do and do not qualify as correct responses. Fluency can be quantified by counting the number of correct responses which can be produced in a given period of time (e.g. correct responses per minute), or by measuring the length of time which is required in order to perform a particular

behaviour, or to complete a particular task to a given standard.
In other words, measurement of all of the behavioural dimensions identified in the preceding section can be accomplished using just two measurement operations: (a) counting and (b) timing.
The observation, quantification and recording of time poses no particular difficulties. There are well established procedures for measuring elapsed time and well established procedures for checking the accuracy of measures of elapsed time. Units of time (seconds, minutes, hours, and so on) have agreed definitions and the definitions are independent of particular timing devices. Units of time are standard units and time scales are equal-interval scales with a true zero. This means that measures of time in terms of seconds, minutes and hours can be manipulated arithmetically and can be compared against each other.
The kind of scale which is produced by counting particular responses (e.g. correct responses) depends upon the nature of the response which has been selected for counting. If each response (in the set of responses which has been selected for counting) is a response which involves a similar amount of effort for its performance and which can be performed by a competent person in a similar amount of time then increments in the count will tend to produce an equal-interval scale with a true zero. That is, a count of 1 response will function as a standard unit upon which arithmetical operations can be performed.
If there is some doubt as to whether or not each of the responses in a set of responses are equivalent in this way, an empirical check is possible. Someone who can perform all of the responses in a competent fashion can be asked to perform all of the responses in the set (or a representative sample of them), the time taken to perform the responses can be divided into a number of equal intervals (e.g. ten, 30-second intervals), and the number of responses which are performed during each interval can be checked to ensure that closely similar numbers of responses have been performed

during each of the intervals. If it is found that the responses which have been selected for counting are not equivalent (as might be the case with the response "writing answers"), then a smaller and more circumscribed response may be selected for counting (e.g. "writing words").
Sensitivity
A measurement procedure may be accurate over a given range but the range may be too restricted to cope with real-life variations in the behavioural dimension which is being measured. In other words, measures of behaviour change must be sensitive as well as accurate measures of change.
"Sensitivity refers to the capability of measured variation in the dimensional quantity to reveal changes in the phenomenon of interest to the investigator" (Johnston and Pennypacker, 1980, p141). In order to ensure that a given measurement procedure has an adequate degree of sensitivity, a number of conditions must be met.
First, the observation period must be long enough for changes in performance to be detected. Let us say, for example, that the observational period has been set at one hour, that the behaviour of interest is "asking for help", and that the actual rate of occurrence of

the behaviour is once every two hours. If, during an observational period, the behaviour is not observed, it cannot be concluded that the rate of occurrence is zero, but only that it is less than once per hour. The measurement procedure which has been selected (observing for one hour) is not sufficiently sensitive to detect the exact rate of occurrence, nor is it sufficiently sensitive to detect changes in rate as long as the true value remains below one per hour.

This problem is solved by changing the measurement procedure. In this case, by increasing the length of the observational period until an accurate estimate of the true frequency of occurrence of the behaviour can be obtained. In the above example, this might be accomplished by accumulating observations over, say, ten 1-hour observation periods. Needless to say, the choice of an appropriate measurement procedure (e.g. the choice of an appropriate observational period) requires some prior knowledge of the current frequency of occurrence of the behaviour which is to be recorded (Yarrow and Waxler, 1979).

A similar problem arises when the measuring instrument is a test. If the test is to detect changes in performance, the test must make provision for a sufficient number of response opportunities on the part of the learner. Let us say, for example, that improvements in the ability to apply some problem solving procedure are being measured with a test which contains just two test items. (This is not a ridiculous example. Many standardized achievement tests contain only one or two items which test particular skills). The only scores which are possible on a 2-item test are 0, 1 (50%) or 2 (100%). A measurement scale which has only three values clearly cannot provide a sensitive measure of improvements in the skill which is being tested.

This problem is also solved by changing the measurement procedure - in this case, by increasing the number of items in the test of competence. The exact number of response opportunities which the test should provide depends upon the degree of sensitivity which is desired. White and Haring (1980) report that the introduction of daily measures of performance on their own tend to produce improvements of the order of 3% per day. Given this observation, it follows that a daily test of, say fluency, should contain a minimum of 33 response opportunities since this is the number of responses which are required in order to detect the effects of daily measurement alone. Note that the number of response opportunities may need to be much greater than 33. For example, some instructional decisions (such as the decision to promote to the next reading level) depend on the observation of a low error rate (such as a reading error rate of less than 5%). In a case such as this, the total number of response opportunities (e.g. words to be read) should be in the region of 100 plus, rather than in the region of 33 plus.

Secondly, the measurement procedure must be one which does not artificially restrict the degree of improvement which the student can demonstrate. The following examples illustrate this requirement.

If the administration of a test is artificially paced as, for example, when items are carried on flash cards and the flash cards are presented by the

teacher or experimenter, then the maximum level of fluency which can be demonstrated by the learner is the level at which flash card items are being presented. A young child who has mastered letter names, for example, may be able to name letters at a rate in excess of 60 letters per minute. If the letters are on flash cards and the cards are being presented at the rate of 1 every 3 seconds, the highest level of fluency which the child can demonstrate is 20 letters per minute. In this particular example, the testing procedure is insensitive to changes in fluency in excess of 20 per minute because pacing imposes an artificial ceiling at 20 per minute. Here again, the problem is avoided by changing the measurement procedure - perhaps by placing all of the flashcards in front of the learner prior to testing.

A similar problem arises when too much time is allowed for a test of competency. Let us say, for the purposes of example, that a daily test of multiplication facts contains an appropriate number of items (30) and that the time allowed is three minutes. A child who has mastered her tables may be able to answer these items at a rate in excess of 30 items per minute. If a 3-minute test is used, the child always has three minutes and the highest fluency rate which can be demonstrated is 30/3 or 10 per minute. This test is not sensitive to improvements in fluency in excess of 10 answers per minute. This problem can be avoided either by measuring the time taken to complete 30 items, or by reducing the time available, or by increasing the number of items in the test until it exceeds the number which can be completed by even the most competent child.

A similar kind of problem arises when the measurement procedure measures only changes in the accuracy with which the student can respond. Learners often achieve 100 per cent accuracy long before they achieve skilled, fluent, or automatic performance. If the teaching aim is mastery or fluency then the only appropriate measure of learning is one which is sensitive to changes in both accuracy and speed of performance (Duncan, 1974; Howell and Lorson-Howell, 1990; Lindsley, 1990; West, Young and Spooner, 1990).

Thirdly, the content of the measure of learning must match the content of the instruction which is being evaluated. The most sensitive measure of instructional effect is the test which examines only the skill (or skills) which a given sequence of instruction has been designed to teach and which the learner has yet to learn. If the test includes items which test skills which the student has already mastered prior to instruction, no improvement can be observed on these items. If the test contains items which test skills which were never covered during instuction, then no improvement can be observed on these items. A test which mixes items which test both taught and untaught skills provides a much less sensitive test of instructional effects than a test which consists only of items which test taught skills. In traditional measurement theory, such a test would be said to be lacking in "content validity".

## Part 3
## Examples of continuous measures of learning

Attempts to develop continuous measures of learning which meet the

technical requirements set out in the preceding section result in testing procedures which are very different to the testing procedures which are used to measure pupil achievement in the traditional learning experiment. These differences are well illustrated by projects which have been undertaken by education students enrolled in behaviour analysis courses at the University of Canterbury. Education students at Canterbury study behaviour analysis at the Stage 2 level (where the focus is on basic principles of behaviour), at the Stage 3 level (where the focus is on acquisition) and at the M.Ed. level (where the focus is on the management of behaviour problems and learning difficulties in schools). Regardless of the course or level, all students in behaviour analysis courses are required, as part of their course requirements, to complete a course of practical work and to design and carry out a short experimental analysis of behaviour in a human subject. A more detailed description of these courses will be found in Church (1992a). The studies described below were completed by students in the third year course (EDUC 324: Individual Learning Processes) and the graduate course (EDUC 650: Behaviour Management).

Figure 1 (adapted from Murphy, 1988) presents the results of an experimental study of the development of touch typing skills in a 12-year old girl. To measure the child's progress, Murphy asked the child to type a test sentence at the end of each day's lesson. The test sentences, which were different each day, were matched for difficulty. Tests were limited to 3 minutes. Performance on the test was converted to the number of characters correctly typed per minute. Rate of improvement was limited by the teaching procedure which introduced new keys at the rate of two keys per day. The daily lessons, which lasted for about 20 minutes, consisted primarily of practice on the new keys and on keys already introduced on previous days. No difficulties were experienced in achieving an accurate classification of responses as correct or incorrect and the measure was found to be sensitive to a change in teaching procedure (in this case, withdrawal of feedback about performance on the daily test).

Figure 2 (from McIlhone, 1987) shows improvement in printing skills in a 6-year old boy. Printing skill was measured from day to day using acetate overlays to assess the accuracy with which the child had copied the letters in a 6- to 8-word test sentence. The same set of eight, matched, test sentences were used during both the first and second experimental treatments to control for test difficulty. While the coding of responses as "correctly" and "incorrectly" formed proved to be a challenging task, satisfactory levels of reliability were obtained. As can be seen from the figure, the measure proved to be sensitive to a change in teaching procedure (in this case, introduction of feedback in the form of the opportunity to watch the letters being marked by the teacher).

Figure 3 (from Oxnam, 1991) describes the progress of a young girl aged 4

years 6 months while learning to name the letters of the alphabet. The daily test consisted of a chart containing all of the lower case and upper case letters of the alphabet arranged in random order. At the end of each lesson, the child was asked to "point to and name the letters which you know". Responding was limited to 1 minute. The figure presents the number of letters correctly named per minute from day to day. The measure proved to be sensitive to a change in teaching procedure (practising with letters only versus practising with cards which contained both a letter and an alphabet picture).

Figure 4 (from Davis, 1991) shows the results of a 7-year old boy's attempts to learn to tell the time, to the nearest quarter hour, from an analogue clock face. To measure progress, the teacher prepared 48 pictures of clocks showing each of the 48 times to be learned. These were set out in random order on six sheets with eight clocks on each sheet. At the end of each lesson the child was given the six

sheets in a random order and asked to point to and say the times which he knew, working through the sheets at his own pace. Responding was limited to 2 minutes. In this study, the author has presented both the correct responses and the incorrect responses made by the pupil. Skips were not recorded.

Figure 5 (from Alderston, 1991) presents the results of 24 days remedial tuition in basic multiplication facts with a 13-year old boy. The daily test consisted of 50 randomly ordered multiplication questions set out in horizontal form. The pupil was asked to complete as many questions as possible in a 1 minute period. The test was taken at the end of each day's lesson. The figure shows the number of digits correctly and incorrectly written during these tests (for example, if the pupil responded to the question "3 x 4 = ___ " with the answer "12", this was counted as 2 digits

correct). Lessons included verbal and written practice with flash cards and sheets of written problems. The baseline phase shows the rate of improvement which resulted from the daily testing on its own. The results of the experiment indicate that the measurement procedure was sensitive to the effects of a change in lesson content (from practice on facts with factors up to 4 to practice on facts with factors up to 7). No difficulties were experienced in accurately classifying digits as correct or incorrect.

Figure 6 (from Whyte, 1991) shows the effect of tuition and practice in colour names with a 28-month old girl. The daily test consisted of a sample of 20 common household items drawn sequentially from a total pool of

320 items. The 20 items contained either three or four examples of each of the following colours: black, brown, orange, pink, purple and grey objects. (The child was already able to identify the colour of red, blue, green, yellow and white objects). A total cumulative latency of 1 minute (3 seconds per answer) was allowed during each test. The accurate classification of responses as correct or incorrect required prior agreement that approximations to the correct pronunciation would be counted as correct. Separate sets of coloured objects were used for teaching and for testing. In this study, the dimension of performance which is being measured is generalization. Each day, the coloured objects which made up the test were completely new objects which had never before been used in the daily test (or in the daily teaching sessions). Examination of the child's performance on Days 15 and 16 reveal a clear ceiling effect with the child providing 20 out of 20 correct responses. This ceiling effect illustrates the problem which occurs when improvement is limited by the testing procedure.

Figure 7 (adapted from Sandford, 1991) shows the results of 32 days of oral reading practice by a 13-year old boy with a reading age of approximately 8 years. Daily tests consisted of a 1 minute oral reading probe taken from the current instructional reading book. The figure shows both the number of words read correctly and the number of words read incorrectly during each probe. Reading responses were classified as correct or incorrect using the running record conventions described by Clay (1972), that is, self-corrections were classified as corrects and omisions and insertions were counted as errors. All lessons included practice in reading aloud. In addition, the baseline lessons included shared reading and written exercises, the second phase included practice with unknown sight words, and the third, fourth and fifth phases involved preferred activity reinforcement for improvements in speed. The figure shows that the measurement procedure was sensitive to changes in the difficulty level of the reading material. Given this demonstration, it may be concluded that the teaching of unknown sight words had no effect on reading fluency.

Figure 8 (adapted from Rae, 1988) shows the rate of acquisition of new spelling responses in a 9-year old girl. Teaching sessions, which were 20 minutes long, began with a test of the 10 spelling words studied the previous day, followed by 15 minutes practice on 10 new words drawn from a pool of words which the student had been unable to spell during pretesting. Daily sets of words were matched on word length and difficulty. Words were studied under two conditions, first, using the child's previously learned practice strategy and, secondly, using the "Spell-Write" rehearsal procedure (Croft, 1983). The "Spell-Write" procedure was taught between Day 5 and Day 6. The figure shows the cumulative number of new spelling words mastered from day to day. No difficulties were experienced in

classifying responses as correct or incorrect and it is clear from the figure that the measurement procedure was sensitive to the effects of the change in the practice procedure.

The effects of teaching the "Spell-Write" rehearsal strategy have been measured by five separate studies involving seven separate learners and the results of all of these studies have been closely similar (Church, 1990b). In these five studies, rate of acquisition has been measured by counting the number of unknown spelling words mastered from day to day. However, alternative measures of spelling progress are possible. For example, Innes (1978) measured improvement in spelling by counting the number of words correctly spelled during twice-weekly, 100-word, spelling dictations. The difficulty of the dictations was controlled by drawing them from Listener editorials.

Figure 9 (from Osborn, 1992) presents the results from a study of improvements in knowledge. The learner was a 20-year old, female, university student who was studying German History and Culture. The topics studied during the course of this experiment were the Weimar Republic, the Third Reich and the Post War Years. The information to be learned was set out in a study guide supplied by the German Department. The topics in this study guide were first divided into two sets and 50 short answer questions constructed to test recall of the information provided in the study guide about each of the two sets of topics. The daily tests were produced by shuffling the Set A and Set B question cards and drawing off the top 15 questions from each of the shuffled sets. The student had 5 minutes in which to write the answers to as many of the 30 questions as possible. During the student selected study phase the student studied the course materials, covered phrases of her own choice and attempted to write out the phrase which had been covered. During the teacher-selected phase the teacher covered up what appeared to be the most important or relevant words in the study material and the student attempted to supply the missing words. The results of this experiment are shown in Figure 9 which shows the number of correctly answers questions on each daily test. The coding of answers as correct or incorrect was relatively straight forward as most of the correct answers were given in words similar to those used in the study materials.

EDUC 324 students have undertaken a number of similar studies - examining improvements in "knowledge" about penguins, about marine mammals, about dinosaurs, about the weather and so on. Most of these studies suffer from the same shortcoming - the student designing the study has underestimated the rate at which primary aged children are able to acquire new "knowledge" responses with the result that the learner's progress has been severely limited by the low rate at which new items of information have been introduced by the teacher. The Osborn experiment avoids this shortcoming.

Part 4
Discussion

The learning experiments discussed in the preceding section demonstrate that it is possible to develop continous measures of learning which meet accuracy, reproducibility and sensitivity criteria. These examples also illustrate the kinds of information about learning, and the kinds of information about the effects of teaching variables on learning, which can be obtained when the experimenter chooses to use continuous rather than "one shot" measures of learning.

There appear to be few limitations on the kinds of learning which can be studied using continuous measures. The behaviour analysis literature includes examples of the continuous measurement of improvements in a wide range of skills including self-help skills (e.g. Schuster, Gast, Wolery and Guiltinian, 1988), athletic skills (e.g. Shapiro and Shapiro, 1985), social skills (e.g. Bornstein, Bach, McFall and Friman, 1980) and language skills (e.g. Olswang, Bain, Dunn and Cooper, 1983). Students at Canterbury have developed continuous measures of improvement in ball handling skills, gymnastic skills, self-defence skills, getting dressed, bathing a baby, cooking a meal, riding a horse, driving a car, administering an injection, touch typing, operating a word processor, finding a book in a library, playing tunes on various musical instruments, handwriting, punctuating sentences, spelling, reading, translating French, naming colours, playing chess, writing chemical equations, solving quadratic equations, solving trigonometry problems and knowledge of a variety of topics. Humans acquire many different skills during the course of their lifetime, all are appropriate targets of scientific analysis, and there is no scientific reason for limiting the particular skills which might be studied.

The continuous measures used in these demonstration experiments provide a much more sensitive measure than the "statistically significant treatment effect" generated by a between-groups measure of experimental effect. One of the problems with the "significant treatment effect" as a measure of experimental effect is that it is a measure which has only two values. It cannot be used, therefore, to compare the relative rates of learning which result from experimental manipulations which have greater and lesser effects on rate of learning.

The training experiments reported in Section 3 demonstrate that it is possible to develop standard units of measurement for measuring and reporting the degree and/or rate of learning under observation. In the spelling experiment 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. These counts were then converted to a rate (number of new spelling responses acquired per 15 minute training session). This measurement operation produces an equal interval scale with a true zero. The measurement units have additive properties. A rate of 8 words per session represents twice the rate of

learning as a rate of 4 words per session. The measurement operation yields a measurement result which can be subjected to arithmetical operations and it is one in which the rates obtained from one experiment can be directly compared with the rates obtained in other experiments using the same measurement operation (see Church, 1990).

Repeated testing tends to produce improvements in performance (e.g. Alderston, 1991). The fact that a testing procedure is producing improvements in performance is never of concern to teachers (because the aim of teaching is to facilitate improvement), but it may be of concern to an experimenter. If the aim of the daily tests is to measure the effects of differing instructional procedures, then some time must be spent measuring the effect on performance of simply completing a succession of daily tests. This needs to be done before the experiment proper commences. Provided the same testing procedure is used throughout the experiment, continuous measures of learning do not function as a confounding variable since the effect of the measurement operation is the same from one phase of the experiment to the next.

A test which is given only once may be used to measure achievement but it cannot be used to measure learning. A one shot measure of achievement, no matter how carefully it is developed, item analysed, standardized and marked can never provide an adequate measure of learning because it cannot measure performance change. This is true regardless of whether the measure of achievement is norm-referenced or criterion-referenced. (Where the aim is to measure learning, the debate between norm-referenced and criterion referenced testing is irrelevant.)

Continuous measures of learning, on the other hand, may be used to measure both learning and achievement. This they do much more economically and more accurately than is the case with the traditional achievement test. A succession of daily 1-minute probes takes much less time to design, administer and mark than is the case with the traditional type of achievement test. The results of continuous testing produce more accurate judgements regarding achievement level because such judgements are based upon a number of performance samples rather than upon a single performance sample.

A test which is given only once may be used to make placement decisions, but it cannot be used to make instructional decisions. A one shot measure of achievement, no matter how skillfully designed, cannot provide information about the rate of improvement which is occurring during a particular sequence of instruction.

Continuous measures of learning, on the other hand, do provide information about rate of change during instruction and can therefore be used as a basis for ongoing decisions regarding teaching procedures during instruction. If a daily test shows that a pupil is not improving, this state of affairs becomes apparent within two or three days and the teacher can change the teaching procedure before the pupil experiences weeks or months of failure.

Research into the psychometric properties of continuous measures of

learning indicates that they are as reliable or more reliable than one shot measures of achievement and that their predictive validity is as good or better than that of one shot measures of achievement. "Probe data are as psychometrically sound as standardized achievement tests, are much simpler to administer, and are much less expensive. In addition they are repeatable and thus may serve a monitoring function for the efficacy of instruction" (Lentz, 1988, p98).

References

Alderston, S. (1991). Effects of precision teaching techniques on multiplication facts proficiency of a 13-year old boy. Unpublished EDUC 650 Case Study report. Education Department, University of Canterbury.
Bornstein, P.H., Bach, P.J., McFall, M.E., Friman, P.C. and Lyons, P.D. (1980). Application of a social skills training program in the modification of interpersonal deficits among retarded adults: A clinical replication. Journal of Applied Behavior Analysis, 13, 171-176.
Church, R.J. (1990). The use of within-subject 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 the effects of teaching on 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, Department of Education.
Clay, M.M. (1972). Reading: The patterning of complex behaviour. Auckland, Heinemann.
Cone, J.D. (1981). Psychometric considerations. In M. Hersen and A.S. Bellack (Eds.), Behavioral assessment: A practical handbook (2nd ed.), New York: Pergamon Press.
Crossman, E.R.F.W. (1959). A theory of the acquisition of speed-skill. Ergonomics, 2, 153-166.
Croft, C.O. (1983). Teachers manual for Spell-Write. Wellington, N.Z. Council for Educational Research.

Davis, M.J. (1991). The effects of student and random choice of instructional material on learning to tell the time using the analogue method. Unpublished EDUC 324 Field Study report. University of Canterbury, Education Department.
Duncan, A.D. (1974). Tracking behavioral growth: Day-to-day measures of frequency over domains of performance. Educational Technology, June, 54-59.
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.
Hays, W.L. (1967). Quantification in psychology. Belmont, Brooks/Cole Publishing Company.
Howell, K.W. and Lorson-Howell, K.A. (1990). Fluency in the classroom. Teaching Exceptional Children, 22(3), 20-23.
Innes, A. (1978). Effects of sixteen weeks tutoring on the spelling ability of a university student with a spelling error rate of one in thirteen words. Educational Research Newsletter, No 10, 24-27.
Johnston, J.M. and Pennypacker, H.S. (1980). Strategies and tactics of human behavioral research. Hillsdale, Lawrence Erlbaum Associates.
Lentz, F.E. (1988). Direct observation and measurement of academic skills: A conceptual review. In E.S. Shapiro and T.R. Kratochwill (Eds.), Behavioral assessment in schools: Conceptual foundations and practical applications. New York: The Guilford Press.
Lindsley, O.R. (1990). Precision teaching: By teachers for children. Teaching Exceptional Children, 22(3), 10-15
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.

Olswang, L., Bain, B., Dunn, C. and Cooper, J. (1983). The effects of stimulus variation on lexical learning. Journal of Speech and Hearing Disorders, 48, 192-201.
Osborn, A.M. (1992). The effects of requiring relevant and irrelevant overt student responses on the learning of German history facts. Unpublished EDUC 324 Field Study Report. University of Canterbury, Education Department.
Oxnam, J. (1991). The effects of providing meaningful information about letters of the alphabet to a four-and-a-half year old. Unpublished EDUC 324 Field Study Report. University of Canterbury, Education Department.
Pennypacker, H.S. (1976). Measurement, accountability, and the economics of a complex instructional system. In L.E. Fraley and E.A. Vargas (Eds.), Proceedings of the Third National Conference on Behavior Research and Technology in Higher Education. Society for Behavioral and Technology Engineering.
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.
Sandford, C. (1991). Effects of contingent access to a menu of free time activities on the reading rate of two Third Form boys. Unpublished EDUC 650 Case Study report. University of Canterbury, Education Department.
Schuster, J.W., Gast, D.L., Wolery, M. and Guiltinian, S. (1988). The effectiveness of a constant time-delay procedure to teach chained responses to adolescents with mental retardation. Journal of Applied Behavior

Analysis, 21, 169-178.
Shapiro, E.S. and Shapiro, S. (1985). Behavioral coaching in the development of skills in track. Behavior Modification, 9, 211-224.
Simkins, L.D. (1969). The basis of psychology as a behavioral science. Waltham, Blaisdell Publishing Co.
West, R.P., Young, K.R. and Spooner, F. (1990). Precision teaching: An introduction. Teaching Exceptional Children, 22(3), 4-9.
White, O.R. and Haring, N.G. (1980). Exceptional teaching (2nd ed.). Columbus, Charles E. Merrill Publishing Co.
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.
Whyte, L.A. (1991). The effect of overt practice on the correct naming of the colours of common household objects by a young child. Unpublished EDUC 324 Field Study report. University of Canterbury, Education Department.
Yarrow, M.R. and Waxler, C.Z. (1979). Observing interaction: A confrontation with methodology. In R.B. Cairns (Ed.), The analysis of social interaction: Methods, issues and illustrations. Hillsdale, Erlbaum.