INSTRUCTIONAL VARIABLES AND LEARNING OUTCOMES

Robert M. Gagné

University of California, Berkeley

From the Proceedings of the

SYMPOSIUM ON PROBLEMS IN THE EVALUATION OF INSTRUCTION

University of California, Los Angeles
December, 1967

M. C. Wittrock, Chairman

Sponsored by the Center for the
Study of Evaluation of Instructional Programs

The research and development reported herein was performed pursuant to a contract with the United States Department of Health, Education, and Welfare, Office of Education under the provisions of the Cooperative Research Program.

CSEIP Occasional Report No. 16, September 1968
University of California, Los Angeles

INSTRUCTIONAL VARIABLES AND LEARNING OUTCOMES

Robert M. Gagne

Evaluation is a word commonly used to refer to a great variety of activities connected with educational programs. In its broadest sense, it has to do with valuing, or determining the worth of, educational courses, programs, or even whole systems. It has, however, a considerably more modest meaning when applied to the appraisal of a very important, although sometimes small, segment of a total program or system, namely, of the extent to which specified instances of learning have occurred.

Although one can conceive of "evaluating" learning outcomes in something like a cost-effectiveness sense, this is not usually what one wishes to do. Instead, interest often centers upon the accomplishment of certain human performance objectives as a part of a set of more comprehensive goals. For example, a learning outcome pertaining to a student's mastery of differential equations may be merely a portion of a larger goal of producing a capable mechanical engineer. When considered in relation to this kind of specific learning outcome, it may be that evaluation is better expressed as <u>assessment</u>. The latter word may imply the desirable characteristic of objectivity (as opposed to the subjective nature of "valuing") and at the same time carry an implication of the importance of such an activity to the larger evaluation goal.

So long as one looks upon education, or perhaps only schooling, as a system having a definable social purpose, whose functions and components are subject to planning and design, assessment of learning outcomes may be seen to have an essential importance within the system. One can, of course, study separately the characteristics of certain other parts of the system, such as the method of communicating, the subsystem of guidance, or the functions of the teacher. One can even study separately the processes which take place during system operation, such as student-teacher interactions or teacher-administrator interactions. All of these are useful to know about. But, so far as I can see, nothing can take the place of the student's performance as an absolutely essential criterion of system or subsystem functioning. There may be many reasons to know how teachers are conducting their questioning, how administrators react to an innovation, or whether students enjoy going to class. But none of these can take the place of learning outcomes as an essential part of any seriously purposed "evaluation" of educational systems or subsystems. By definition, the effecting of externally stimulated behavioral change in students is a major purpose of education, and this implies that behavior assessment must be undertaken.

Obviously, one can assess outcomes soon after the occurrence of some educational processing, or at a later time, even considerably later. There is often a need to discover, for example, not simply whether a method of solving differential equations was mastered immediately following a period of time devoted to its learning, but also whether it is remembered several weeks later

when the student is faced with learning a more complex method incorporating the formerly learned technique as one step. Or again, one may be interested in whether, at a later time, the method of solving differential equations can be recalled and used in connection with quite a different situation, like that involving the rotary motion of a body or the rate of a chemical reaction. The question of assessing learning outcomes evidently must include those human performances affected by the processes of retention and of transfer of learning.

The Problem of Measurement

Scientific measurement is generally agreed to be fundamentally a matter of counting units which are agreed upon as being generated by the same operations. As the writings of Campbell (1957) indicate, the operations applied to counting the measures referred to as length, mass, and time have been fairly easy to agree upon, and, therefore, may perhaps deserve to be called "fundamental." Most measurement in science, however, is not of this fundamental sort, but instead is derived from it. For example, the chemist considers that he has a satisfactory measure, let us say, of the "strength of a solution" when he performs those operations which extract the solid components and then relates the mass of these components to the total mass of the solution. Or the biologist may construct an indirect measure of the "adiposity" of tissue by ascertaining the proportion of area in cells observed under a microscope taken up by material which is stained a particular color. In such instances, and there are many of them, the

scientist is constructing a measure of some entity by demonstrating its relationship to an operation of counting.

It is apparent that what is done in measurement, so far as intellectual operations of a measurer are concerned, is that an inference is made. The chemist infers the variable "strength of solution," the biologist, the variable "adiposity." These variables are not directly observed, as for example, color, shape, and numerosity may be; each is an inference depending upon a chain of reasoning. In the same manner, the psychologist infers the variables of learning, retention, and transfer of learning. None of these variables is observed directly. Each of them is measured indirectly by being related rationally to operations that include counting.

One of the major implications of this line of reasoning is the following: in undertaking measurement, one must be prepared to answer two questions, one of which precedes the other. The first question is, "What is being measured?" There must be, in other words, agreement among those who use measurement that the units defined by a set of operations are the same and, therefore, can be given a common name. One measures length by counting units which can be matched and thus readily agreed upon as being the same. But indirect measurement, as in the case of measures of "solution strength," "adiposity," or "learning," may offer a much more difficult problem of agreement, because a more complex set of operations is involved. Accordingly, the answer to the question "What is being measured?" is often not an easy one to determine. It seems probable to me, for example, that virtually

all controversy in the field of learning research over the past several decades could be categorized in the question "What is being measured?".

The second question of measurement is "how much?". When measurement is indirect this too is not always an easy question to answer--even when the operations are agreed upon, because of the problem of size of unit. If the unit, for example, is an apple, there are problems about what will be agreed upon as a "standard apple," how one will handle variations in the size of apples in applying such a unit, and so forth. But even when such measurement problems are encountered, they fade to insignificance beside the monumental confusions in measurement which derive from mixing apples and bananas. This is why the question "What is being measured?" is the first question, and in that sense the more important one. To worry about the scaling of units for a mixture of apples and bananas is really quite ridiculous. The first problem is one of demonstrating agreement that one indeed has something called "apples" and something called "bananas" to measure.

Indirect measurement, then, requires that there be a defined set of operations as a basis for agreement on the inference as to what is measured. Before one worries about how much, it is necessary to distinguish the measurement operations for one inferred entity from those for another, as well as from what may be called the "noise" in the system as a whole. This thought may be summed up by saying that two primary criteria of measurement are (a) distinctiveness and (b) freedom from distortion. These two criteria need to be applied if we are to have measurement at all.

Learning, Retention, Transfer

For three rather gross classes of behavioral inferences, it should be possible to apply these primary criteria to the question of what is measured. It would not seem to be too difficult a task to show that the categories called learning, retention, and transfer of learning can be examined by means of these criteria. And if this is so, perhaps finer inferential categories (such as varieties of learning) can be similarly examined.

Learning. The inference called learning appears to depend upon the following set of operations. First, it is determined that an individual cannot do a particular performance A. (The operations used to make this inference are themselves specifiable, as will become apparent.) Second, the individual is provided with a certain sequence of stimulation, and it is determined that he is attending to this stimulation. Again, the operations required to demonstrate attention are simply mentioned here, but are capable or being specified. Third, another set of operations is used to determine that he is motivated to perform. Finally, the observation is made within a specified brief time that the individual either does or does not exhibit performance A. To be complete, one also makes the observation that he exhibits performance A', another member of the class A (and perhaps also that he exhibits performance A''). As a result of these last observations, it would be generally agreed that the inference is justified that the individual possesses performance capability a. In other words, his nervous system, at least during this brief time, has a capability which makes possible peformances of the class A.

Has capability a been learned? This is the inference sought. In order to make it, however, one must carry out still another operation on the same kind of individual by repeating the total procedure but omitting the second step, the sequence of external stimulation. This control makes it possible to infer learning rather than growth. Of course, this operation is often assumed rather than actually carried out, as is true with some of the other steps. When the measurer is satisfied on this point, he may then make the inference that capability a has been learned.

Distortion of measurement is avoided in these operations by procedures used to insure that attention is in effect, that there is motivation to perform, and that more than a "chance" performance has been observed. The inference of learning is not justified unless these means are taken to insure that the measurement is free from distortion. If one is concerned with "How much?" obviously the factors of motivation, attention, and variability of response can affect the measurement. But more important is the fact that distortion can reduce the amount to zero, and thus have a direct affect on the question "What is measured?".

Distinctiveness of measurement pertains in this instance to the distinction between learning and growth as justifiable inferences. It is noteworthy that the demonstration of distinctiveness in this case ideally requires a "control" operation, which in some circumstances becomes a control group. A large performance change takes place over a relatively small, arbitrarily chosen period of time, which in most cases can be agreed upon as not produced by growth. For this reason the control operation is frequently not

actually carried out. Nevertheless, it is of considerable importance to recognize that it is rationally demanded, and cannot be ignored. Its assumption needs to be explicit.

Retention. Retention is inferred from a set of operations somewhat as follows. First, there is a measure of what may be called "immediate learning effects," taken in accordance with the procedures previously described, within a specified time--usually a few minutes following the application of the stimulus situation for learning. (As studies of "short-term memory" make us aware, it is important that the time for this measure be set at a few minutes rather than at a few seconds; further discussion of this point, however, will not be undertaken here.) Here is another instance in which it is necessary to make the measurement of immediate learning outcome in a separate equivalent subject or group, in order that the measure of retention remain uncontaminated. There are scores of studies in the older literature which suffer from this methodological defect in the measurement of retention.

After it has been demonstrated that there has been some immediate effect of learning, the measurement of retention may be undertaken at some specified time--hours, days, or months after the learning session has been completed. Again, interest centers upon the inference that capability a has been retained, as shown by the execution of performance A. For such an inference to be valid, the performance measured after the intervening time must, of course, be the same as that measured "immediately."

The inference about retention of capability *a*, however, is not quite this simple, and usually certain other precautions of measurement must be observed. This is because what happens to the learner during the intervening time has some marked consequences. As a single example, it is known that the learner may engage in "internal rehearsal"; in fact, it is difficult to prevent him from doing so (Murdock, 1963). In addition, it is known that different kinds of intervening activity, introduced with the purpose of preventing such rehearsal, have different effects on retention as finally measured (Loess & McBurney, 1965). At the very least, it may be said that a measure of retention is uninterpretable without a specification of what has happened to the learner in the period between learning and the measurement of retention.

Again in this instance of measurement, one can see the need to apply the criteria of freedom from distortion, of distinctiveness, in order to be sure about what is measured--before one faces the question "How much?". Distortion is prevented by operations which control the opportunities for "internal rehearsal," and which also control the kind of intervening activity known to produce varying amounts of interference (cf. Postman, 1961). Distinctiveness is insured by control-group operations which make possible the inference that that capability, which has been learned in the first place, has (or has not) been retained.

Transfer of Learning. Measuring transfer involves many of the operations previously mentioned, and others besides. The inference one wishes to make is that some capability *a*, exhibited in performance A, having been learned and retained, has an effect

on the learning of capability b (in some performance B'). The capabilities a and b are different in some respects that are specifiable. It is evident that the inference of transfer depends upon measurement operations that involve the demonstration that learning and retention are present as prior events. Otherwise, of course, one may not know whether what is hypothesized to transfer is present in the first place. Transfer is also markedly subject to contamination by intervening events, as is the case with the measurement of retention. Studies of retroactive inhibition (cf. Keppel, 1968) provide the classical setting for the masses of evidence bearing upon this measurement problem.

In this case, too, the question of what is measured is subject to the criteria of freedom from distortion and distinctiveness. For the former, one must design operations to demonstrate that learning has occurred, that retention is possible, and to control and specify intervening events. Distinctiveness of measurement must be insured, as was true with learning and retention, by the use of a control subject or group which demonstrates the absence of effects from sources other than the capability on which interest centers. One must bear in mind that one wants to make the inference that capability a has transferred, not simply that capability b has been facilitated or interferred with.

Classes of Learning Outcomes

It appears to be so, then, that the distinguishing of learning outcomes into the gross categories of learning, retention, and transfer requires measurement operations which are designed

to distinguish "place" learning from "sequence" learning. Hunter's (1913, 1920) studies of delayed reaction and double alternation were designed to distinguish between the learning of discriminations and "representative process." Harlow's (1949) work on monkeys proposed to draw a distinction between "discrimination habits" and "learning sets." Investigators of the learning of verbal sequences are concerned with devising measures which will distinguish the learning of sequence from the learning of item position (Jensen & Rohwer, 1965). Many other examples could be given.

In my work (Gagne, 1965), I have suggested eight different kinds of inferences which seem to me to be generally useful distinctions to make throughout the field of learning as a whole, particularly as they are relevant to school learning. I am prepared to think that more than eight distinctions may be important to make, and that there may be several reasons for making them. Nevertheless, it still seems to me that these eight categories are of particular significance to learning research and theory, as well as to the particular affairs of education.

In repeating these categories here, I do not wish to cover old ground. Rather, I should like to examine specifically the question of distinctiveness between pairs of these categories. As is true with other classes of learning outcomes, each of these inferred capabilities carries its own set of problems with respect to the criterion of distortion (that is, effects of other variables on performance). But it would be too long a job to consider these distortion effects here. The question of distinctiveness, however, seems to be of particular relevance to a consideration of "what is

measured" when interest centers upon "what kinds of capabilities can be inferred?"

The distinctions among inferred capabilities which should be described for present purposes are as follows:

1. The classical conditioned response ("signal learning") versus the operant conditioned response (response learning").
2. The operant conditioned response versus the motor chain; or versus the verbal sequence.
3. The single response or chain versus the multiple discrimination.
4. The multiple discrimination versus the (nonabstract) concept.
5. The concept versus the principle.
6. The concrete principle versus the abstract (or higher-order) principle.

Two Types of Conditioning. The problem of distinguishing the measurement operations for classical and operant conditioning has a long history, extending back to Skinner's (1937) landmark paper, and even before that. Modern writers (e.g., Kimble, 1961; Grant, 1964) state the major distinctive operations to be somewhat as follows: In the classical conditioning situation, the conditioned response being observed is maintained by the pairing of conditioned and unconditioned stimuli, where such pairing is independent of the learner's response; in operant conditioning, the maintenance of the learned response depends on presentation of the "unconditioned stimulus" in a manner which is contingent upon the occurrence

of that response. What is most notable about this distinction, for present purposes, is that it takes at least two observations to establish distinctive measurement. One must know, first, that the learned response does depend upon the presentation of a conditioned stimulus, and second, that the learned response either is or is not maintained when the contingency of conditioned response followed by unconditional stimulus does not obtain.

Single Connection and Chains. The question of distinctiveness also arises in connection with the learning of a single connection versus the learning of chains. In motor learning, such a demonstration could presumably be carried out by showing that a set of describably different responses forming a sequence (such as unlocking a lock with a key) can each be initiated by a separate stimulus. If each element or link in the chain can thus be separately demonstrated, then each may be contrasted with the total sequence as a single connection. Of course, one may continue the measurement process by seeing whether each "single" connection may, in its turn, be further broken down into two or more additional links. Presumably, there is at least a practical limit as to how far this process may be carried.

Another important example of the distinctiveness criterion occurs in the learning of verbal paired associates (verbal chains). Whereas for many years the paired associate was treated as a single connection, the studies of Underwood and Schulz (1960) have demonstrated the necessity of observing at least two links in a chain. Thus, operations were devised to measure "response learning" separately from "association learning," and no modern

investigator would think of ignoring this distinction. Nowadays, a number of procedures are used to insure that measurement is either directed primarily at the "associative phase," for example, by using response words that are known to be previously well-learned or else at "response learning" itself, by having the learner recall freely a set of originally unfamiliar syllables. Establishing the distinctiveness of single connections and verbal chains is usually either a two-stage measurement process, or one which accomplishes a similar purpose by the use of experimental and control groups.

Single Connections and Multiple Discrimination. Studies of discrimination learning in both animals and human beings usually make the distinction between single connections and multiple discriminations quite clear. Typically, this is done by using the kinds of single connections which are already known to be well-learned, or which can be shown to be. Discrimination in a white rat, for example, can be measured by testing whether he jumps to the right or left, or perhaps presses a lever with a vertical or horizontal push. In either case, one attempts to observe the discrimination only after assuming that the jumping, or the lever-pressing, has been previously learned. When such prior learning of single connections (or chains) cannot be assumed, it must first be demonstrated. An investigator would not try to observe discrimination learning unless the differential responses called for could not be shown separately as being already present in the animal's repertoire.

Lists of verbal associates also fall into the category of multiple discrimination. Each single pair, provided its component

links are previously learned, is retained with near-perfection for a short time following a single presentation (Murdock, 1961). When one or more pairs is added, the increased difficulty in learning that occurs is one of the best-known phenomena in the verbal learning field. Difficulty of learning increases with length of list (McGeoch & Irion, 1952). The effects of intralist similarity of interferences among items have also been extensively studies (cf. Underwood, 1964. Without attempting to describe further the various findings bearing on this point, let me simply say that investigators of paired-associate learning are careful to distinguish the measurement of single verbal associates from the measurement of sets of associates (which I here put in the category of multiple discrimination). The single verbal chain is learned in one trial, and this usually serves as a standard against which to measure the learning of sets of associates that are greater than one.

The Use of Control Techniques

These brief descriptions of measurement techniques applicable to single connections, chains, and multiple discriminations have been included not with the intention of an exhaustive consideration of their measurement problems, but rather to provide a background for discussion of the measurement of more complex learning outcomes. It seems to me that the learning of concepts and principles, the major domain with which school learning is concerned, is likely to face similar measurement problems and be subject to similar criteria, as are these simpler kinds of learning. If there are techniques for applying the criterion of distinctiveness to measuring these simpler kinds of learning outcomes, similar techniques

should at least be tried in the measurement of more complex capabilities.

I refer to concepts and principles as being more complex than connections, chains, and discriminations for the very simple reason that they appear to require the inference, or the postulation, of more elaborate mechanisms to account for them. In other respects, they may not be more complex; for example, the conditions typically required to bring them about by learning may actually be simpler to describe. Whatever the case, it is certainly true that they have been studied less, and one has many fewer pieces of evidence to call upon in identifying appropriate measurement procedures.

What does the scientific literature on simpler learning processes suggest about the measurement of concepts and principles as outcomes of learning? What kind of extrapolations can be made concerning the problem of distinctive measurement, of identifying precisely what is being measured?

The theme of methodology running through the application of measurement to simpler capabilities appears to be this: the criterion of distinctiveness requires that a control be employed before a dependable conclusion can be drawn about what is measured. Sometimes it is possible to make this control measurement on the same person, sequentially with the second measurement which more specifically encompasses the learning outcome of interest. When this is possible, one may think of distinctive measurement as a two-stage process, the first (control) stage of which must be done before a firm conclusion can be drawn from the second stage. In other instances, control takes the form of using another (equivalent)

individual, or another group. In such instances, the two measurement procedures, control and "experimental," can be applied at the same time or within the same experimental context. But in either case, the purpose is the same: it is to measure the capability distinctively; to insure that what one wants to measure is not in fact something else.

Let me state this proposition more specifically. Control procedures of measurement are used to distinguish classical and operant conditioning; if the latter is being measured, one must either first or independently demonstrate that the response being observed does not occur when the contingency of instrumental production of the conditioned stimulus is made impossible. Only then is one justified in being convinced that what is being measured is an operant response. Similarly, control procedures are used to distinguish single connections and chains; one can legitimately speak of chain learning only when independent measurement has shown that the of component links in the chain have been previously learned. Distinctive measurement procedures also apply to single connections and multiple discriminations: one measures discrimination learning only when independent methods have been used to demonstrate that the single connections or chains which make up the multiple set to be learned have already been acquired. (In animal learning studies, the control operation is usually specifically carried out; in human verbal learning, it is usually assumed to require a single trial.)

When one turns to the procedures used in observing concepts and principles as learned outcomes, one is immediately struck by

the fact that control measures are often not employed. Instead, the general picture seems to be dominated by quite a different set of procedures, derived from other kinds of considerations. One begins to encounter measurement instruments described as "multiple-choice tests," "completion questions," or "matching questions." These may well be useful ways to apply measures, and nothing can be immediately perceived as wrong with them. But still another characteristic of measurement emerges in the fact that these techniques are not usually employed in such a way that controls are present. The single "item" of a single type appears to be employed as the unit of measurement, rather than a two-stage technique or a control procedure. To be sure, more than one item is usually employed, but the justification for this is reliability (a variety of the freedom from distortion criterion), as opposed to distinctiveness of measurement. Logically, no amount of concern with reliability can provide a solution to the problem of identifying what is measured.

There is, then, an apparent discrepancy which needs to be further examined. The kinds of learning outcomes most frequently measured in laboratory studies are usually subject to the criterion of measurement distinctiveness. This requires some kind of control technique to insure the distinction between one kind of learning outcome and another. In contrast, the kinds of learning outcomes most frequently encountered in educational settings are typically measured with techniques that do not involve control procedures. Is it possible that such procedures are unnecessary with these more complex kinds of learning? Can suitable

assumptions be made to make them unnecessary? Or has their applicability been somehow overlooked?

Measuring Concepts and Principles

It is now time to examine more closely the kinds of distinctiveness considerations that may be applicable to the measurement of concepts and principles, and in addition, their desirability as criteria of measurement. In doing this, I shall discuss the distinction between multiple discrimination and concepts, concepts and principles, and among different classes of principles.

Multiple Discriminations and Concepts

In dealing with concepts, as a first step it is desirable to define the term "concept." By a concept I mean the kind of capability that enables an individual to identify (by class name or otherwise) a specific member of a class of objects, object properties, actions, or events, when that specific member is new to him. This is the kind of entity the psychologist studies, usually in children, when he deals with the learning of colors, shapes, textures, positions, and directions. A somewhat special category of such concepts, with which I shall not deal further here, are the kinds of concepts that have designated combining rules, such as "conjunctive concepts" (yellow and bordered), "disjunctive concepts" (either yellow or bordered), and the like (cf. Bruner, Goodnow, and Austin, 1956). In still another and quite different category are concepts that cannot be conveyed by giving experience with specific members of a class, but which must be communicated by <u>definition</u> (cf. Gagne, 1966).

The simpler kinds of concepts, having concrete referents, are what I have in mind to discuss here. Defined concepts, which appear to be formally identical with principles, will be dealt with later.

In order to make an observation to detect whether a concept has been learned, one must in effect ask the question: "Among this variety of objects (object properties, actions, events) that I show you, which are bilpads?". (The final word represents a name for the class.) I say "in effect," because it seems evident that such a question can functionally be asked of an animal who does not understand language, as was true of Harlow's (1949) monkeys, who learned to respond to concepts such as those we call "right," "left," and "odd."

However, it is at once apparent that asking a question like "Which are bilpads?" is in itself an insufficient condition for making a distinctive measurement of the concept. Suppose that "bilpad" means a particular shade of tan, and that objects having a somewhat different shade of tan are not "bilpad." The individual who has not previously learned to discriminate these shades of tan will not make correct identifying responses and, thus, will not have learned the concept according to the measurement applied. A recent analysis by Martin (1967) relates the concept to multiple discriminations in terms of S-R connections. While Martin's hypothesis of mutual inhibition is not essential to the conception of concept here being discussed, his clear exposition of the necessity of discrimination as a prior condition of concept acquisition is highly relevant.

The purpose of distinctive measurement of a concept capability is, after all, to assess the effectiveness of some particular learning situation. Suppose that the learners have been placed in this learning situation, designed to have them acquire some designated concept. When measurement operations are applied, it is found that some have acquired the concept, while others have not. But of those who have not, there are two kinds, and a single-stage sort of measurement will not be able to distinguish them. Specifically, there will be learners who (a) have previously acquired the necessary discriminations, but who do not know the concept, and learners who (b) have not previously acquired the necessary discriminations, and who do not know the concept.

Single-stage measurement, of the sort which simply tests whether or not learners correctly identify "bilpad," or some other concept, is not distinctive measurement. Such single-stage operations do not distinguish concepts from multiple discriminations or, more specifically, the absence of concepts from the absence of multiple discriminations. The design of distinctive measurement seems a relatively simple matter, but it must involve a control which determines the presence or absence of concepts. A two-stage operation would probably be simplest to use--one which first measured the presence of multiple discriminations, and then the presence or absence of the concept.

I see no reason why this line of reasoning does not apply directly to the measurement of concepts that are learned in school, whether they are very simple ones such as the printed letters learned in the early grades, or more complex ones like "cell nucleus" learned in some later grade. Again, the criterion of distinctive measurement requires that a distinction to be drawn

between those performances which fail to identify members of the class "cell nucleus" when the learners can make the necessary discriminations, and those performances which fail to identify the concept when the learners cannot make these discriminations. Two-stage measurement would appear to provide a way of achieving such distinctiveness.

Concepts and Principles

I should now like to consider the measurement of principles, or ideas, as contained in such simple statements as "leaves grow on trees," "airplanes fly in the sky," or "a nervous impulse is a wave of electrical dipolarization propagated along the membrane of a neuron." It is evident that measurement of such principles also presents very similar problems so far as the criterion of distinctiveness is concerned.

Each principle is composed of concepts. The principle "leaves grow on trees" is, very simply, composed of the four concepts, "leaves," "grow," "on," and "trees." Since concepts are involved, a single-stage measurement operation simply does not provide distinctive measurement, because the failure of a learner to demonstrate that he "knows the principle" may mean that he does not "know the concepts." The principle that "a parallelepiped is a prism whose bases are parallelograms" may obviously not have been learned because the learner does not have one or more of its component concepts, whether "prism," "base," or "parallelogram." Distinctive measurement of what is learned would seem to require two-stage measurement, the first stage devoted to assessing whether the concepts have been acquired, and the second to whether the principle is known.

Still another problem of distinctive measurement occurs with principles. This is the problem of distinguishing the measurement of principles, not from concepts, but from verbal associates. This problem occurs particularly because of the fact that verbal items are typically and widely used to measure the acquisition of principles. Accordingly, the problem is to distinguish knowing the principle from knowing the names of the concepts which make up the principle. Obviously, a test item like the following,

Leaves grow on_____,

or any variants of this item using multiple-choice alternatives, may be measuring the verbal association "leaves-trees" rather than the principle itself.

Ideally, this difficulty would be overcome by using actual members of the classes of concepts involved (that is, real trees, real growing, and real leaves) to make up the stimulus situation in which the measurement is taken. Employing a two-stage process, the concepts themselves would first be identified by the learner, who would then be asked to demonstrate the principle. Under certain other assumptions, distinctive measurement of a principle can be carried out by asking the learner to demonstrate by means of a picture. For example, the instruction may be given to a child (following upon a first stage of concept measurement): "You have learned that leaves grow on trees. Draw a picture to show this."

The criterion of distinctiveness in the measurement of principles has often not been followed in experimental studies of substance learning. In their review of investigations thirty

years ago, Welborn and English (1937) mention several widely used methods of measurement, including (a) verbatim recall with verbatim scoring; (b) free recall with scoring for main ideas; and (c) recognition measures employing multiple-choice tests. More recent examples of each of these methods can also be readily located, such as Newman (1939), Ausubel, Robbins, & Blake (1957), and King & Russell (1966).

It appears evident at once that the verbatim method of scoring fails to distinguish between the retention of verbal chains and the retention of principles. The employment of control procedures to measure these two outcomes separately is deliberately undertaken in some studies (e.g., Cofer, 1951; King & Russell, 1966), and these procedures surely acknowledge the problem. It is not quite so evident, however, that all investigators who have used this kind of control are entirely clear about what it is they want to measure. Even less useful, from the standpoint of distinctive measurement, is the method of recognition using multiple-choice tests. Such items do not require the recall of a principle, since the major portion of the principle is included either in the stem of the item or in one of the alternatives. If such a partial statement of the principle repeats the learning statement verbatim, we again face the problem of indistinguishability from verbal sequence learning. On the other hand, if the partial representation of the principle contained in the test item is a paraphrase, the recognition of the missing word may be a matter of identifying a concept, rather than a reinstating of the principle.

I am led to the belief, therefore, that measuring the learning outcome of principles must be accomplished in some way or other that requires the learner to demonstrate the idea of the principle, and that this must be distinguished by suitable control procedures from both verbatim learning of a verbal sequence and from the learning of the concepts which make up the principle. One method, already suggested, would require identification of the concepts as a first stage, followed by some sort of representation (as in drawing) of the principle as a second. Assuming that a suitable first stage is employed, the second stage might well take other forms, one of which is paraphrasing (as used, for example, by English, Welborn, & Killian, 1934).

Another instance of failure to describe and exemplify distinctive measurement of principles occurs in the work of Bloom and his collaborators (1956), in their treatment of a taxonomy of learning tasks. These authors equate the category of "knowledge" with "recall of ideas," and illustrate their measurement mainly with multiple-choice test items. One simply does not know what these items measure. Here is an example (Bloom, 1956, p. 81):

"Magnetic poles are usually named:

1. plus and minus.
2. red and blue.
3. east and west.
4. north and south.
5. anode and cathode."

An item such as this might be measuring the verbal association "magnetic pole--north, south." Many people would probably say that

is what it does measure. If so, then it can hardly be considered to constitute adequate measurement of a principle. One can imagine an item in this general category designed to measure a concept, as in the following instance:

(Picture of a magnet, with lines of force)

Label the poles of the magnet by their usual names.

However, it is apparently not possible in this instance to turn such an item to the purpose of principle measurement, because it simply does not imply a principle. One can imagine an item designed to measure such a principle as the following: When two bar magnets are allowed to attract each other, the north pole of one will be adjacent to the south pole of the other. The item would be:

(Picture of two bar magnets)

The two bar magnets in the picture have come together by attraction. Label the poles of each.

These examples illustrate one way in which two-stage measurement might be employed. First, a measure is applied to determine whether the individual possesses the concept of magnetic poles. If he knows this, it is then possible as a second step to attempt to test whether or not he has learned a principle. Obviously, if the second step is undertaken without the first, the results are ambiguous. They do not enable us to distinguish learners who have not learned the concept "magnetic pole" from learners who have learned this concept, but who have not learned the principle.

Distinctions Among Principles

In the field of achievement testing, it is not difficult to find examples of measurement which do not meet the criterion of distinctiveness. Often these instances fail to distinguish a simpler, more concrete principle from a more abstract one. More generally, they attempt to measure more than one principle at one and the same time. An example I have used before is the following (Gagne, 1965, p. 259):

> (Diagram of a pipe showing a cross section)
> Find the thickness of a pipe whose inner circumference is 9π, and whose outer circumference is 21π.
> (a) 12π; (b) 12; (c) 6π; (d) 6; (e) 3.

This item fails to distinguish measurement applicable to two different principles. First is the principle relating circumference to diameter, i.e., $C=\pi d$. Second is the principle, rather readily applicable to the accompanying picture, that the outer diameter of a pipe equals the inner diameter plus twice the thickness of the pipe.

Clearly, if one wishes to know whether these two principles have been learned, they must be measured separately. Again in this instance a two-stage measurement operation, or some other form of control, is called for. If the individual fails this item, we do not know whether (a) he did not know the first principle; (b) he did not know the second principle; (c) he did not know either principle; or (d) he knew both principles, but was unable to put them together in solving a new problem. The criterion of distinctiveness in measurement is not met by items such as this, which try to combine too many measurement operations into one.

Requirements of Learning Measurement

The substance of what I am saying is this: Measurement of learning outcomes in laboratory studies of such varieties of learning as connections, chains, and multiple discriminations has generally been characterized by careful attention to what is being measured. For those varieties of learning in which such analysis and definition of outcome has not been true in the past, one can readily see pronounced trends in this direction at present. Measurement of these simpler sorts of learning events appears to be subject to the criteria of (a) distinctiveness, by which is meant operations that distinguish one class of learning outcome from another, and (b) freedom from distortion, involving a set of operations which distinguish learning from the action of other variables of various sorts.

The criteria of distinctiveness and freedom from distortion, when applied to these relatively simple kinds of learning, appear to require the use of control operations as essential parts of the measurement itself. Often, this means the employment of control individuals or groups. In other instances, control is exercised by using the same individuals, but making the necessary operations in two stages. The first stage is the control measurement, and the second is the measurement which is the center of interest.

When one looks at more complex varieties of learning outcome, it appears that control operations are equally applicable and equally necessary, if one is going to be confident about what is measured. Thus, it is possible to design control operations for

measurement distinctiveness that make possible differentiation between concepts and multiple discriminations, between principles and concepts, between principles and verbal sequences, and between principles of differing levels of abstractness. Although not elaborated here, it is not difficult to believe that operations to avoid distortion of measurement are similarly feasible for these varieties of learning.

Such measurement operations for concepts and principles are possible; however, they appear not to have been much used. Instead, there is what I would characterize as a regression to the use of inexact techniques originally designed for quite different purposes. These techniques are those of mental testing, which were designed primarily for the purpose of predicting performance, rather than of measuring learning outcomes. If one is concerned with prediction, it is likely to make little difference whether a distinction can be drawn between the learning of, say, a concept, and the learning of a principle. In brief, the individual who "knows more" is going to exhibit faster learning and a better ultimate performance, regardless of what the particular components of his capability are.

It should be clear that the purposes of achievement measurement in the schools are not always those of prediction. In many important instances, the purpose is to identify what has been learned, and by so doing to relate it to that which was intended to be taught. Thoughtful writers on achievement testing have usually given considerable emphasis to this distinction. Lindquist (1951), for example, states that the first step in designing

achievement tests must be that of determining what is to be measured. Nevertheless, having acknowledged this idea, most writers on achievement measurement proceed blithely to describe techniques of test design which pay no further heed to the methods of distinctive and distortion-free measurement.

Dependence upon imprecise techniques of mental testing for the measurement of concepts and principles has led to the ignoring of the requirements for control procedures which have come to be standard features of measurement techniques used with other simpler kinds of learning outcomes. The single item used in traditional achievement testing constitutes an uncontrolled, ambiguous measure which can only in rare instances be shown to be related directly to the learning outcome of interest. Partly as a consequence, perhaps, there is appeal to "scores" from a set of items which must be shown to have elaborate statistical relationships to each other. Statistical methods are used in the attempt to increase precision of measurement, when in fact the problem of achieving distinctiveness and freedom from distortion could best be accomplished by direct control procedures.

What is suggested by this review is that techniques of control need to be developed and used for the measurement of concepts and principles as outcomes of learning. In particular, the notion of two-stage measurement, rather than dependence on the single item (alone or in collections), seems to offer greatest promise of a feasible solution. Of course, if this suggestion were followed, it would lead to new kinds of tests for achievement, as well as

new kinds of scoring procedures. Neither of these are now in existence, but they appear to be technically possible.

I need to mention again the question of measurement of "How much?". There are some highly interesting problems involved in this question, but I cannot begin to deal with them here. Their neglect should not be interpreted as indicating, however, that I consider them unimportant. Instead, I have simply followed the advice of other investigators of measurement, and attempted to give priority to the question of what is measured. If this priority matter can be clarified, it should not be difficult to devise clear definitions of what may be meant by degree or amount of learning.

Designing techniques to measure learning outcomes seems to me to be a most important requirement for research on complex forms of learning, and consequently also for the practice of educational measurement. We need some new procedures, based upon the criteria of distinctiveness and freedom from distortion, to accomplish the measurement of concepts and principles. These procedures should take into account the use of controls which characterizes the experimental measurement of such learning outcomes as connections, motor and verbal chains, and multiple discriminations in laboratory studies. In this way, a consistent logic of measurement may ultimately come to pervade the entire field.

REFERENCES

Ausubel, D. P., Robbins, L. C., & Blake, E., Jr., Retroactive inhibition and facilitation in the learning of school materials. Journal of Educational Psychology, 1957, 48, 334-343.

Bloom B. S. (Ed.). Taxonomy of educational objectives. Handbook I. Cognitive domain. New York: David McKay, 1956.

Bruner, J. S., Goodnow, J. J., and Austin, G. A. A study of thinking. New York: Wiley, 1956.

Campbell, N. R. Foundations of science. New York: Dover Publications, 1957.

Cofer, C. N. A comparison of logical and verbatim learning of prose passages of different lengths. American Journal of Psychology, 1941, 54, 1-20.

English, H. B., Welborn, E. L., & Killian, C. D. Studies in substance memorization. Journal of General Psychology, 1934, 11, 233-259.

Gagné, R. M. The conditions of learning. New York: Holt, Rinehart, & Winston, 1965.

Gagné, R. M. The learning of principles. In H. J. Klausmeier & C. W. Harris (Eds.), Analyses of concept learning. New York: Academic Press, 1966. Pp. 81-95.

Grant, D. A. Classical and operant conditioning. In A. W. Melton (Ed.), Categories of human learning. New York: Academic Press, 1964. Pp. 3-31.

Harlow, H. F. The formation of learning sets. Psychological Review, 1949, 56, 51-65.

Hunter, W. S. The delayed reaction in animals and children. Behavior Monographs, 1913, 2, 1-86.

Hunter, W. S. The temporal maze and kinesthetic sensory processes in the white rat. Psychology, 1920, 2, 1-17.

Jensen, A. R., & Rohwer, W. D., Jr. What is learned in serial learning? Journal of Verbal Learning and Verbal Behavior, 1965, 4, 62-72.

Keppel, G. Retroactive and proactive inhibition. In T. R. Dixon & D. L. Horton (Eds.), Verbal behavior and general behavior theory. Englewood Cliffs, N. J.: Prentice-Hall, 1968.

Kimble, G. A. Hilgard and Marquis' Conditioning and learning. New York: Appleton-Century-Crofts, 1961.

King, D. J., & Russell, G. W. A comparison of rote and meaningful learning of connected meaningful material. Journal of Verbal Learning and Verbal Behavior, 1966, 5, 478-483.

Lindquist, E. F. Preliminary considerations in objective test construction. In E. F. Lindquist (Ed.), Educational measurement. Washington, D. C.: American Council on Education, 1951, Pp. 119-158.

Loess, H., & McBurney, J. Short-term memory and retention-interval activity. Proceedings of the 73rd Annual Convention of the American Psychological Association, 1965, 85-86.

Martin, E. Formation of concepts. In B. Kleinmuntz (Ed.), Concepts and the structure of memory. New York: Wiley, 1967. Pp. 33-67.

McGeoch, J. A., & Irion, A. L. The psychology of human learning. New York: Longmans, Green, 1952.

Murdock, B. B., Jr. The retention of individual items. Journal of Experimental Psychology, 1961, 62, 618-625.

Murdock, B. B., Jr. Short-term retention of single paired associates. Journal of Experimental Psychology, 1963, 65, 433-443.

Newman, E. B. Forgetting of meaningful material during sleep and waking. American Journal of Psychology, 1939, 52, 65-71.

Postman, L. The present status of interference theory. In C. N. Cofer (Ed.), Verbal learning and verbal behavior. New York: McGraw-Hill, 1961. Pp. 152-179.

Skinner, B. F. Two types of conditioned reflex and a pseudo type. Journal of General Psychology, 1935, 12, 66-77.

Underwood, B. J. The representativeness of rote verbal learning. In A. W. Melton (Ed.), Categories of human learning. New York: Academic Press, 1964, Pp. 48-78.

Underwood, B. J., & Schulz, R. W. Meaningfulness and verbal learning. Philadelphia: Lippincott, 1960.

Welborn, E. L., & English, H. B. Logical learning and retention: A general review of experiments with meaningful verbal materials. Psychological Bulletin, 1937, 34, 1-20.