DOCUMENT RESUME

ED 158 649 HE 010 345

AUTHOR Bligh, Donald
TITLE A Pilot Experiment to Test the Relative Effectiveness of Three Kinds of Teaching Method and Gropings for a Design of Objective Tests of the Effectiveness of Teaching Methods.
INSTITUTION London Univ. (England). Inst. of Education.
PUB DATE 71
NOTE 22p.
AVAILABLE FROM University of London Institute of Education, UTMU, 55 Gordon Square, London WC1H ONT, England

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *Cognitive Processes; *Educational Objectives; Educational Research; *Effective Teaching; Foreign Countries; Higher Education; Independent Reading; *Learning Processes; Lecture; Logic; Methods Research; *Multiple Choice Tests; Reading Materials; Student Evaluation; Tape Recordings; *Teaching Methods
IDENTIFIERS Cognitive Taxonomy of Educational Objectives; Great Britain

ABSTRACT

Two papers on teaching methods are presented. The first concerns an experiment that tested the relative effectiveness of three teaching methods: an uninterrupted lecture, a tape-recording, and reading a prescribed text. The three-way research design used the three teaching methods, three groups of six student teachers of physical medicine between the ages of 24 and 42, and three psychological topics as subject matter. Each teaching method was followed at once by a multiple choice test designed to measure eight cognitive levels from Bloom's "Cognitive Taxonomy of Educational Objectives." These levels of thinking are: terminology, facts, generalizations, understanding, application, analysis, synthesis, and evaluation. A discussion of these concepts, testing conditions, and results of the pilot experiment is provided. The second baper concerns objective tests of the effectiveness of teaching methods in higher education. The correspondence between logical processes in the student's mind and categories of Bloom's "Cognitive Taxonomy of Educational Objectives" are expressed in terms of propcsitional, or truth functional, logic. The analysis concerns testing: a student's knowledge of fact presented during instruction, knowledge of presented relations betweeen facts, unpresented relations between presented facts, the ability to apply presented principles and generalizations, the ability to analyze unstated generalizations, unpresented relations between presented facts, the ability to evaluate, and value judgments. (SW)

UTMU
University Teaching Methods Unit

University Teaching Methods Unit
University of London Institute of Education
55 Gordon Square London WC1H 0NT

A PILOT EXPERIMENT TO TEST THE RELATIVE EFFECTIVENESS OF THREE KINDS OF TEACHING METHOD

and

GROPINGS FOR A DESIGN OF OBJECTIVE TESTS OF THE EFFECTIVENESS OF TEACHING METHODS

by

DONALD BLIGH

PRICE 30P

A PILOT EXPERIMENT TO TEST THE RELATIVE EFFECTIVENESS OF THREE KINDS OF TEACHING METHOD

Most of the time a student spends at his work involves receiving information either by reading or by listening to lectures; but which is the more effective in terms of the objectives of further education?

Reading has the advantage that the student may go at his own pace, think about what he is doing and go over it again if he does not understand. It is sometimes claimed that the lecture is effective because the presence of the lecturer is stimulating and he in turn responds to the reactions of his audience. A tape-recording played to a class seems to have the disadvantage of both, it is impersonal and paced.

According to Bloom (1) the objectives of receiving information form a hierarchy, from knowledge of its terminology to its evaluation, and students may be tested at these different levels by multiple choice questions (M C Q).

The aim of this experiment was to reach a provisional conclusion as to the relative effectiveness of an uninterrupted lecture (L), a tape-recording (T) and reading a prescribed text (R), in terms of the levels of thinking outlined in Bloom's "Cognitive Taxonomy of Educational Objectives" measured by multiple choice questions.

The design was a three way latin square using three teaching methods (L, T, R), three groups of six student teachers of physical medicine between the ages of 24 and 42 (A, B, C) and three psychological topics as subject matter (S, G, M).

		L	B	R - Teaching Method
Sets	A	M	G	S
	B	G	S	M
	C	S	M	G

Subject Matter

The procedure was such that for each topic (S, G, M) a lecture was given to one group of students without interruption and lasted for 20 to 30 minutes. The lectures were recorded and played to a second group of students without the teacher being present. Care was taken that where the lecturer (L) had used the blackboard either for illustration, or for emphasis of certain words, the same stimuli were available in condition T.

A typist duplicated the lectures using a dictaphone machine. Punctuation was at her discretion. Fundamental illustrations were also duplicated, and words that were written on the blackboard were underlined.

Each teaching method (L, T, R) was followed at once (allowing time to give out question and answer papers) by a multiple choice test at the following eight cognitive levels:

Table 1.

Level	1.1 Terminology	ie recognition of a name or term
"	1.2 Facts	ie recognition or recall of specific facts
"	1.3 Generalisations and Principles	
"	2.0 Understanding	ie simple interpretation and relations of facts
"	3.0 Application	ie facts and principles
"	4.0 Analysis	ie of data given in the question or in S,G or M
"	5.0 Synthesis	including a plan proposal or creative act
"	6.0 Evaluation	judgements about facts in S, G or M

The students who were taking a fairly intensive course in psychology for two weeks before this experiment, were relatively "naive" psychologists. Hence, although one cannot be certain, it seems reasonable to suppose that the thinking required to answer the higher level questions was largely done while receiving the information or during the tests, and not before. This point is important if one is to assert that the differences in cognitive level are attributable to the methods of receiving information. To avoid the criticism that the thinking at higher levels may have been stimulated by questions earlier in the test, questions at different levels were placed in a random order.

All the multiple choice questions were five-choice situations and were untimed. There are two controls in the setting of questions which are very difficult to observe, but which may have a common solution: (a) devising distractors, which are not only of equal probability with others in the same question, but are comparable with those in other questions in such a way that the differences in difficulty between questions of different cognitive levels are derived from the differences in mental processes required, rather than the fineness of discriminating the correct answer from its alternatives; (b) constructing questions at a particular cognitive level on Bloom's scale. Bloom admits that (b) is difficult, but it seems to be further complicated by apparent sub-hierarchies of concepts within Bloom's. This may be seen from the following groups of words associated with Skinner's "instrumental conditioning":

(A) Food, rat, lever, milk, ...

(B) Stimulus, response, motivation, reinforcement, ...

(C) Rate of response, conditioning, shaping, extinction, spontaneous recovery, ...

Superficially it would seem that the meaning of all these words could be tested at level 1.11, "Terminology"; but those in (A) are ordinary words in everyday language. The teacher may assume them to be known before he does any teaching; indeed he must teach by using them. Those in (B) are technical terms capable of definition in terms of (A) or by use of examples described in (A). Being technical terms they must be taught by the teacher. Words at level (C) are defined in the technical terms already defined by the teacher, i.e. level (B). Thus the student must have more than common knowledge to learn words at level (C). Furthermore words at level (C) usually involve combining concepts from the previous level for their definition. This will require an understanding of those concepts and understanding is at level 2.00 on Bloom's scale. Therefore in this experiment questions testing the meaning of terms contained in (C) were regarded as of the same level of difficulty as "understanding" words at level (B). Words at level (A) were not tested.

Therefore it is suggested that problem (b) above can be made easier if all the words and concepts of a subject to be taught are allocated to groups (A)...(N) according to their complexity before the tests are devised. Problem (a) can be made easier if we adopt a rule of thumb that all distractors in a question should be from the same word group or conceptual level as the correct answer. Thus the degree of discrimination required of the student is the finest possible. This also ensures that the student goes through all the cognitive processes analysed by Bloom to obtain the right answer.

To some extent Bloom anticipates the possibility of there being sub-hierarchies, and his categories allow for this. Yet he cannot know how many there may be because this may vary from one piece of subject matter to another. Since a name can be given to almost anything, sub-hierarchies could start at almost any Bloom level, especially 4.00 and 5.00 (Analysis and Synthesis), not only level 2.00.

The results of a 3-way analysis of variance of the subject matter topics, sets of students and teaching methods each showed insignificant variations when total scores were taken into account. When these totals were broken down into the eight cognitive levels shown in table 1 interactions with subject matter and sets of students remained insignificant, but the interaction between cognitive level and teaching methods was significant at the 1% level. Inspection of Figure 1 shows that students did better at levels 5.00 and 6.00 after tape-recordings than after reading or listening to lectures. There was no significant difference between the teaching methods at the simpler cognitive levels.

As should be predicted Figure 1 also shows that students score higher at the simpler cognitive levels than at the more complex levels. The differences between total scores at the various cognitive levels would occur by chance less than 1 in 1,000 times.

Comments. In view of the disadvantages of tape-recordings mentioned at the beginning, their superiority was wholly unexpected by the experimenter. It might be argued that the novelty of this method of learning in the classroom, plus the need to concentrate in case essential facts were missed, led to greater self-discipline; but the results do not support this interpretation which would require superior retention and understanding of facts, (i.e. Bloom level 1.00 - 2.00) following tape-recordings. Could it be that auditory presentation of verbal information is superior to a written presentation, and that the presence of a lecturer is actually distracting rather than stimulating? If so, a cassette library with play-back facilities using ear-phones should be seriously considered by college librarians. Such a service could be available during normal hours, it could save teachers time, it is suitable for revision especially for weak students, it is cheap, and if used by individuals it is self-paced which is an advantage for foreign students. However, the results of this experiment and the effectiveness of individual use both require further empirical verifications.

It will be noticed from the graph that there is a sharp drop in scores between levels 3.00 and 4.00. It is interesting that in their previous training in physical medicine these students were required to learn terminology, facts, how to deal with facts, and how to understand and apply them; but they were not required to analyse or resynthesise them in new ways, still less to evaluate them in relation to other fields of knowledge. (See table 1). The prediction that students from other backgrounds would show a more gradual decline in scores with cognitive level has subsequently been confirmed with student teachers. This suggests that cognitive skills can be taught.

The results of this pilot test sugest that, provided the rule of thumb is adopted, multiple-choice tests using Bloom's Taxonomy are a useful tool for measuring cognitive ability and future experimenters on teaching methods would be unwise to neglect the higher cognitive levels if these could reasonably be included amongst the teaching objectives. Judging from the range of incorrect answers selected, the rule of thumb is a useful tool for equalising the power of distractors (problem a).

Criticisms

(i) The multiple-choice tests were all given immediately following the teaching in order to eliminate the effects of subsequent reading and discussion. This is clearly artificial as one of the purposes of teaching in further education is to stimulate discussion and interest that will lead to further study. It is also a dubious measure of the effectiveness of the teaching in terms of "permanent" learning and in terms of Bloom's levels. Immediate testing obviously favours retention and understanding of facts and principles (i.e. levels up to 2.00). Analysis (4.00), Synthesis (5.00) and Evaluation (6.00) take more time. One learns to absorb the facts during a lesson and then go away and think about them. It may be, however, that one would expect level 3.00 (application) to be thought about after a lesson, not during it. Yet performance at this level is virtually as good as the lower levels. This criticism is an inherent problem in doing any assessment of the effectiveness of teaching methods. Without testing very soon after teaching, it is very difficult to allow for the effects of subsequent activity.

(ii) No generalisations should be made as this pilot test was made using only one teacher.

(iii) No measure of personality or intelligence were taken, but it is usually thought these are important in teaching and learning.

(iv) It could be argued that a multiple-choice test is itself an analytical situation in that the student is required to select one item from a possible five. Thus the nature of the test requires an ability at level 4.00. However, it may be replied that most other forms of testing require some ability at level 5.00 because essay writing or formulating a set of words for the purposes of communication is a synthetic task. Thus it may be that this, too, is an inherent difficulty in any test situation.

(v) It is sometimes said that multiple-choice tests are not objective because there is a subjective element in the setting of the tests. It is difficult to see how this can be eliminated and it seems fair to suppose that even if multiple-choice tests are not objective, they are at least more objective than most other forms of cognitive testing. The use of the rule of thumb may be a significant advance in overcoming this objection.

Conclusions. It must be re-emphasised that this was a pilot experiment and that the following conclusions are therefore only provisional.

(1) Bloom's classification of educational objectives can be used as a tool for assessing the relative effectiveness of teaching methods when modified by the rule of thumb.

(2) Investigations on the effectiveness of teaching methods should include measures of higher cognitive levels if these could reasonably be included amongst the teaching objectives.

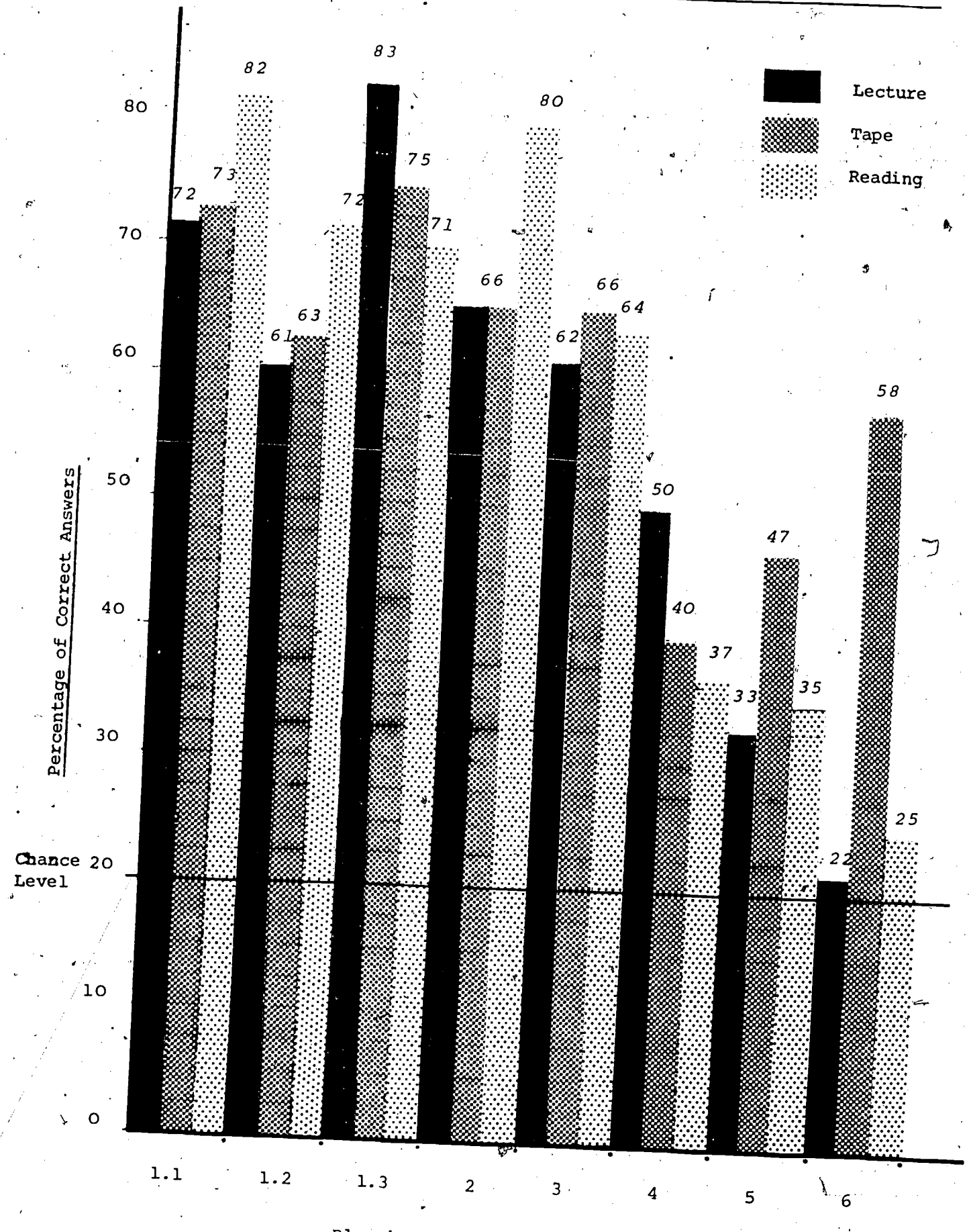
(3) With reservations it seems possible that teaching by use of recorded lectures results in deeper thinking than the use of live lectures and reading prescribed texts; but there is no difference at simpler cognitive levels.

(4) This line of research looks as if it could be promising.

BIBLIOGRAPHY

Bloom B.S. (Ed.) (1954) Taxonomy of Educational Objectives. Cognitive Domain

FIGURE 1 : PERCENTAGE OF CORRECT ANSWERS FOLLOWING LECTURE, TAPE OR READING AT VARIOUS LEVELS OF DIFFICULTY



GROPINGS FOR A DESIGN OF OBJECTIVE TESTS OF THE EFFECTIVENESS OF TEACHING METHODS

There are two common criticisms of objective tests: (1) they can only test factual knowledge, (2) they are only objective in the marking, not in the setting (Head 1968). The erroneousness of the first criticism has become more clear since the publication of sample questions in Bloom's "Taxonomy of Educational Objectives" and a number of workers have used multiple choice questions at various cognitive levels to test the effectiveness of teaching methods in higher education (e.g. Hoare 1970, and Bligh 1970).

The purpose of this paper is to present some gropings towards an answer to the second criticism while maintaining questions requiring a variety of cognitive processes to avoid the first. It is possible that the style of test suggested could be used to test students' value judgements, but this requires further investigation. These gropings are only relevant to tests of the effectiveness of teaching methods administered after a period of tuition in which the statements given to the student are specifiable. These statements will be called "propositions(G)" or "statements(G)".

When propositions(G) are specified it is possible to produce further propositions, which may be called question statements, statements(Q) or propositions(Q), and which bear a precise logical relationship to them. This logical relationship is objective. It is open to public scrutiny and public verification according to the canons of logic which are also public. Propositions(Q) may be true or false and their relationship to propositions(G) may be valid or invalid. They are questions in that, although expressed in the indicative, a student is required to say whether they are TRUE, FALSE or he DOES NOT KNOW. Therefore it is possible to design a situation in which a student will answer correctly if he makes certain logical inferences which are specifiable. The converse does not follow. He may answer correctly without thinking logically either consciously or unconsciously. Nevertheless the probability of his doing so is describable according to statistical "laws" which are equally public, and hence, equally objective. The provision of the "don't know" category may reduce the proportion of "guesses", and the incentive to guess may be further reduced by arrangement of the marking scheme (Crow Goldsmith and Diament 1969).

Since there is an infinite number of possible relationships between propositions(G) and propositions(Q) there is a task of selecting particular relationships for the purpose of testing the effectiveness of teaching. The relationships selected should correspond to the objectives of the teaching. Since this paper is considering objective tests of the effectiveness of teaching methods in general, and since Bloom's is a general statement of objectives, this paper is concerned with logical processes in the mind of the student that correspond to eight of Bloom's categories as set out below. Since it is proposed that students should judge the truth or falsity of propositions and the validity of relations between any propositions it will be most convenient to consider these relations in terms of propositional, or truth functional, logic. Propositions unexpressed during teaching and propositions(U).

Bloom's Categories	True Propositions(Q)
1.1 Knowledge of Terminology	$f(a(G))$
1.12 Knowledge of Specific Facts	$(fa)(G)$
1.3 Knowledge of Generalisations and Principles	$(ga \supset fa)(G)$
2. Comprehension	$r(U)$ where $p(G) \,.\, q(G) \supset r$
3. Application	$(ga' \supset fa')(U)$ where a' is a specific instance of 'a' and 'ga ⊃ fa' is given.
4. Analysis	$(a)(fa \supset ga)(U)$ or $(a)(fa \supset \sim ga)(U)$ or $\sim(a)(fa \supset \sim ga)(U)$ or $\sim(a)(fa \supset ga)(U)$
5. Synthesis	$v(U)$ where $(p(G) \,.\, q(G) \supset r(U)) \,.\, (s(G) \,.\, t(G) \supset u(U)) \,.\, (r \,.\, u \supset v)$
6. Evaluation	$r^1(U)$ or $(s^1 \supset r^1)(U)$ where $(s \supset r)(U) \,.\, (p(G) \,.\, q(G) \supset s^1)$ and where 'r^1' is an instance of 'r' and 's^1' is and instance of 's'

The presentation of Table I may lead to two misunderstandings. There is no suggestion that the correspondence between the logical formulae and Bloom's levels is 1 : 1. Bloom's categories are broader and more blurred at the edges because they are intended to be comprehensive and because they have been formed by a concensus of subjective opinion. The logical formulae represent an attempt at precision and objectivity, and are hence restricted.

Secondly although there is increasingly logical complexity it by no means follows that there is increasingly psychological difficulty. An item may be logically more complex if it has more parts; it is at least arguable that some thought processes operate with wholes. (This misunderstanding may arise through using the word "simple" instead of "easy" as the opposite of "difficult". There is also a confusion between the logical description of thought processes and their logical validity.) The degree of difficulty of the questions is an empirical, not a logical, issue and is measured by the formula:

$$\text{Degree of difficulty} = \frac{\text{Total number of correct answers}}{\text{Maximum possible number of correct answers}}$$

The complexity of thought processes is one factor affecting their difficulty. There are many others such as the familiarity of concepts and what they are concepts of.

TESTING TERMINOLOGY

In the symbolism of the predicate calculus we may write a simple statement or proposition (p, q, r, s, etc) as "fa" where "a","b", "c" (or "a^1", "a^2", "a^3" ...."a^n") stand for arbitary proper names and "f","g","h" (or "f^1", "f^2", "f^3" .... "f^n") stand for arbitary predicates, such that "fa" means "'a' has the property 'f'". In other words the predicate 'f' describes or qualifies the subject 'a'.

A proposition "fa" has the simplest form there can be (with the exception of existential propositions with which we are not concerned for the purpose of testing). This produces a problem where we wish to test whether we have taught the general meaning of a proper name ('a'). The proposition "fa" will always be a specific use of it. Inevitably "fa" tests the student's knowledge of the truth or

falsity of a proposition, not a proper name. The degree of generality of "f" can be varied and this problem can be partially solved by making it very general indeed, provided it is not so general that the question can be correctly answered in a pre-test. If "f" is made specific, it is the proposition "fa" that is judged "True" or "False". In other words what is tested becomes knowledge of a fact (albeit the fact that a term is defined in a given way) not knowledge of the use of a term. (I assume here that to know the meaning of a term is to know its use, not its definition which is a fact about it. cf. Wittgenstein).

It may be asked, "What is the criterion of generality? Is this not subjective? If there is only a matter of degree between testing terminology and testing facts, what is the objective criterion by which to distinguish them?" It is tempting to reply that the degree of generality of "f" can be objectively measured by the number of proper names ("a^1", "a^2", "a^3" ...) to which it can be applied. Apart from the practical difficulty of such a criterion, this will not do on a number of theoretical grounds. It seems plausible in instances where the answer is true. For example:

if "fa" = (Intelligence)(a) is (a general name for a number of different abilities) (f)

"f" can be narrowed by specifying the abilities and made more general by progressively omitting details about the abilities until

"$f^n a$" = (Intelligence)(a) is (to do with abilities (f^n).

It ceases to be plausible where "fa" is false but could become true if "f" is made more general. (eg If "Intelligence is a name for divergent abilities" is broadened to "Intelligence is a name of some kinds of abilities".) At the boundary between truth and falsity the student requires the precision of a definition, not a knowledge of the general use of the term. Furthermore the <u>way</u>, or component of the predicate, that is selected for broadening is a subjective choice.

However, although I do not know the answer to this objection it is not a serious one in that the use of the truth-functional formulae as such is not challenged. It is simply a difficulty from trying to use the <u>propositional</u> calculus to test something that is not a proposition, but a part of one.

TESTING KNOWLEDGE ON PRESENTED FACTS

To test a student's knowledge of a fact presented during instruction, he is re-presented with the statement. At its simplest this is a recognition situation, but he may judge it true or false as a result of any number of more complex processes using other information. To some extent the possession of information can be tested by a pre-test (although this has its problems), but not entirely. For example a student may not know "fa" in a pre-test, because he did not know the term "a". In the post-test he may know "a" but not "fa", yet infer it from "ga" also given during teaching. To some extent, too, processes alternative to recognition may be excluded by choosing "fa" from propositions(G) so that there are few related propositions of the kind, "ga" and "fb", particularly if "gb" is also explicit. Since "ga" and "fb" have an objective relation to "fa", this preserves objectivity in principle, but within the ordinary language of a teaching situation, sentence structures are complex and hence the questions are difficult to devise. Furthermore if the teacher judges "fa" to be important and therefore worth testing as one of the cognitive objectives, it is most unlikely that this statement will stand in logical isolation amongst all propositions(G). Indeed the relatedness of a proposition may well be a measure of its importance - but that is another question.

A further problem arises when testing this, and all objectives, that propositions(Q) where the correct answer is "false" involve a more complex thought process than if the correct answer is "true". A student cannot "recognise" a proposition(G) as false (unless it was wrongly taught or was explicitly taught as being false). The realisation that a proposition is false requires some analysis of its meaning, which recognition of whether it has been given before does not. This is particularly true when propositions(Q) are in a list in which most of the questions do not pose a recognition problem, so that the situation facing the student is not a simple question : "Do I, or do I not, recognise this proposition?". In the case of testing factual knowledge this problem cannot be solved by changing the nature of false propositions(Q). It can be partially overcome by ceasing to insist that the testing of factual knowledge should be a precise recognition situation. For example, certain words may be omitted from a proposition(G)

so that the resultant proposition(Q) may be either true or false while the degree of change from proposition(G) does not vary according to its truth value; variation is the source of this problem. Because of these changes the student is now required to remember the meaning of proposition(G), not their form of words. More precisely, it is statements(G) that are to be tested, not propositions(G), where "proposition" is taken to mean a form of words, and "statement" what the form the words says. When thinking in terms of Bloom's categories this change is significant because it introduces an element of "Translation" which, as category 2.10, is classified as testing "Comprehension", not testing facts. However the elements may be small and the fact that questions do not fit Bloom's categories does not invalidate the test design.

Within the context of this paper a more serious criticism is that as soon as the test constructor says "I have changed the wording, but I have not changed the meaning", there is a strong suspicion of subjectivity. A concensus of opinion on semantic rating scales (see Remmers 1964) could relive this suspicion. Alternatively the use of propositions and their negatives in paired pilot tests would show whether correct and incorrect statements(Q) are of equal difficulty; and this is not open to the objection of confusing difficulty with complexity because variables affecting difficulty, apart from complexity, are held constant.

TESTING KNOWLEDGE OF PRESENTED RELATIONS BETWEEN FACTS

It is normally an objective of teachers not only to teach specific facts but certain relations between them. The behavioural criterion of this objective is that the student should be able to recognise statements of these relations (not necessarily the precise propositions(G).)

In truth-functional logic the principal ways of relating propositions are by the "logical connectives" of implication, conjunction, disjunction, and equivalence. For practical purposes implication is the most satisfactory to test. To test conjunction is to do little more than test two facts in isolation. The ambiguity in the ordinary language in which the test is written, between the inclusive and exclusive disjunction could produce ambiguous questions. Statements(G) of equivalence are rare in teaching and consequently pose problems for the test constructor. Implication statements(G) are common and relevant because explanation is a very common objective of

teaching. There are many kinds of explanation apart from logical implication, consequently the relations between propositions to be tested should be broadened to include relations of cause and effect, correlations and other implications accepted within the field of discourse. On the other hand for the purposes of testing, these relations must be restricted to those that are made explicit during teaching (ie they must be presented.) Many implications in teaching are contextual. These should not be included; firstly because they are made by the student as a result of a thought process classifiable as "Comprehension", secondly the inference is sometimes more intuitive than capable of logical expression, and consequently the relation of statements(Q) to statements(G) is not objective.

In order to ensure that it is the relation that is tested rather then the truth of the related propositions, proposition(Q) may consist of an inversion of the explanadum and explanans. Emphasis on the implication can be obtained if proposition(Q) begins "The reason why ....."

Testing presented relations between facts has no precise equivalent amongst Bloom's categories. Bloom's "Knowledge of principles and generalisations" is the nearest in that principles are frequently expressed as implications and generalisations may be; but the relations tested here may be quite specific and in this respect be closer to Bloom's 1.12 or 1.2.

TESTING UNPRESENTED RELATIONS BETWEEN PRESENTED FACTS

Not all the relations between facts that a teacher wishes the student to know, can be presented. The student must relate for himself. Consequently a further teaching objective that should be tested is the student's ability to relate facts not explicitly related by propositions(G). This is a measure of a student's "understanding" (and "interpretation").

If two propositions(G) in conjunction imply a third, this may be stated as a proposition(Q) under this heading provided it was not presented during teaching. That is, if "p . q ⊃ r", "r" may be a proposition(Q) if it is also a proposition(U). False conclusions may be invented.

These questions are not easy to construct because the constructor must make sure that "r" cannot be inferred from proposition(G) other than "p" and "q"

if he is to be able to assert that a specific inference has taken place in the mind of the student who obtains a correct answer (other than by chance). The following example shows that the task of specifying these inferences is not always as straight forward as it might be at first seem.

"p" ≡ "Normal children increase in intelligence as they grow older" (G)
"q" ≡ "Sub-normal children do not increase so much" (G)
"r" ≡ "The difference in intelligence between normal and sub-normal children increases as they grow older" (Q) (U)

"r" is supposed to be a valid inference from "p" and "q". Subjectively it seems reasonable to expect students to make this inference, yet if described in terms of logic it is more complex than $(p \cdot q \supset r)$. "So much as normal children" is assumed in "q". It either assumes the proposition (U) that "sub-normal children are less intelligent than normal children" or the ability of the student to analyse the term "sub-normal" to infer it. This proposition is then used. More important, there are assumptions (U) about the concepts "increase" and "difference" which are hard to specify.

TESTING THE ABILITY TO APPLY PRESENTED PRINCIPLES AND GENERALISATIONS

A further objective is that students should be able to apply a principle and recognise particular instances of a presented generalisation. These may not seem like the same thing since a principle may be expressed "if 'p' the 'q'" where 'p' and 'q' are both propositions with subjects and predicates such as 'a' and 'f', while a generalisation is a single proposition of the form "all 'a' is 'f'" where the subject 'a' has more than one instance. But a single proposition of this kind may be re-expressed "If something is 'a' then it is 'f'" (rather freely). More correctly in symbolic logic this may be written $(a)(ga \supset fa)$. Strawson (1952) also shows that the following forms are permissably equivalent. (The presence or absence of an existential commitment is not of importance here unless a question begs this issue).

No 'a' is 'f'	$(a)(ga \supset \sim fa)$
Some 'a' is 'f'	$\sim(a)(ga \supset \sim fa)$
Some 'a' is not 'f'	$\sim(a)(ga \supset fa)$

Thus generalisations may be expressed as implications.

A student's ability to apply a principle or recognise an instance of a generalisation may be tested by presenting an implication statement (Q) in which the subject of the

first proposition is an instance of the principle or generalisation and in which the second draws an appropriate conclusion (or inappropriate conclusion, if the correct answer is to be "False"). This requires first a realisation of the principle or generalisation to be applied, and second, the student's ability to apply it (ie to draw the conclusion). In practice the left-hand side of the implication may not need to be stated at all (eg "Arithmetic requires convergent thinking" where the generalisation is "Convergent thinking is required where there is only one right answer to a problem ") or it may require two propositions (eg "If a child is 6 years old and has an I.Q. of 120, his M.A. is 7.2"). Yet if the form of the propositions(Q) is variable can it be claimed that the mental process required of the student is always the same? This strikes at the basic assumptions behind the claim of objectivity in the test design.

TESTING THE ABILITY TO ANALYSE UNSTATED GENERALISATIONS

We have seen above that the four classical subject-predicate forms of generalisation may be expressed as implications in truth-functional logic (Strawson 1952). Therefore if these generalisations are presented as ordinary language statements(Q), they may be judged "True" or "False" by students. The instances that form the generalisations (whether truthfully or not) should be obtained from the teaching statements(G).

If a generalisation is expressed in the form (a)(fa ⊃ ga), or one of the other three forms given above and in Table I, the student has to recall each instance of "fa" and check that it is also "ga". (eg "Each of Binet's categories of intelligence has the same range of I.Q." where the ranges in I.Q. were given in a number of proposition(G).) This is an analytical task because it includes identifying a quality that is part of a concept.

TESTING UNPRESENTED RELATIONS BETWEEN PRESENTED FACTS

As it stands this heading is inadequate. Clearly one cannot be concerned with <u>any</u> unpresented facts. They must be relevant. They may include one fact that is judged to be common knowledge or that the student should know from previous study, and one, or both, that may be logically derived from propositions(G). The logical formula by which they are derived need not be invariant but the paradigm is (p . q ⊃ r) and it should be simple. The formula for this stage is

(r . u ⊃ v) where "r" and "u" are derived statements(U), not statements(G), and "v" is classifiable as both (U) and (Q).

This kind of question tests synthesis of what is "understood" and is analogous to Bloom's level 5.00. It may be interpreted as testing a synthesis of two statements testable at level 2.00. Factor analysis of test scores should confirm this interpretation and may be used as a check on abuse of the variant derivation of the synthesised statements.

Allowing this variable derivation is a concession to subjectivity owing to the difficulty of producing level 2.00 statements(Q) previously mentioned, and which occurs three times over in the construction of questions at this level.

TESTING THE ABILITY TO EVALUATE

Evaluation consists of drawing a conclusion by using an unstated principle (or generalisation) and a derived fact as an instance of that principle. In other words it is assumed that the student already possesses certain principles or generalisations (eg "Introspective observations are subjective"). As when testing application the statement(Q) is normally an implication, but may appear as a simple proposition if reformed as a generalisation "all 's' are 'r'", or if the left-hand side of the implication is low in information content or is otherwise redundant. In terms of Bloom's levels this kind of question involves a combination of 2.00 and 3.00. That this is so is similarly confirmable by factor analysis of test responses.

TESTING VALUE JUDGEMENTS

Provided they are expressed in the form of an implication or a subject-predicate proposition, there is no reason why both moral and non-moral value judgements should not be tested by this method. (Indeed it is possible that this method offers a more valid measure where affective statements are hidden amongst cognitive ones, than where students are concious that their values are being examined - but this is mere speculation.) Moral judgements may be expressed in the form of an implication (eg "If (the patient is suffering from shock)(p) then (he should be wrapped in a blanket)(q)"). Non-moral judgements may be expressed by using a value, or "emotive", term in in a subject-predicate proposition (eg "Massage is unhygienic").

It is sometimes objected that value judgements are subjective and cannot be tested, but in the present context they are no more subjective than other logical thought processes. (It is true that there are additional logical rules for value judgements, but they do not prevent the use of this test design and will not be discussed here).

A second objection arises from the belief that there could be no objective canon of "value knowledge" corresponding to the verification of empirical knowledge. But in vocational education there are frequently explicit professional ethics which may be tested by assent or denial to a proposition expressed in public language. Furthermore this is no objection within the context of testing teaching methods where this assent constitutes the teaching objectives.

VALIDATION OF THE TEST DESIGN

Since claims have been made for the objectivity of this test design one might expect that there should be some external criterion to validate these claims. It is possible that a student who does well at, say, application of principles will be good at applying other principles. We might also expect that this student would be good at applying the same principles in different conditions. In other words there is some validation if the skills tested are transferable to other subject matter and to other conditions. If the skills do not transfer then the difference may be either in the subject matter or the conditions. In neither case singly is the test invalidated, for the difficulty of the patterns of logical inference may vary with the subject matter, and if they vary with the conditions that is what the test is designed to show. If test results vary with the same subject matter and the same conditions (this includes control for individual differences) the test may be invalidated because it is unreliable. Reliability is a measure of consistency which is the essence of logic. It is not a coincidence that it is only through this concept that the test design can be invalidated and that it is an internal criterion. Empirical tests alone cannot invalidate logic.

It might be thought that the logic could be validated by reference to some external logic of language. If there was such a logic (such as that proposed by the logical atomists) this could in principle be done, but Wittgenstein (1953) has shown that there can be no such logic. This is why this test design

is only suitable for testing the relative effectiveness of teaching methods, the relative abilities of students or other variables within a given situation. It is a relative measure, No external validation is possible with certainty.

It is, however, possible to establish norms so long as the subject matter, categories of propositions(Q), test conditions, teaching methods and groups of students are specified. The lack of certain external validation is not a defect of the test design so long as it is not used for purposes for which it is not intended.

REFERENCES

J.J. Head	Multiple-choice examinations. New Education Feb. 1968.
D.A. Bligh	A pilot experiment to test the relative effectiveness of three kinds of teaching method. Research in Librarianship No. 15 Vol. 3. 1971.
Crow Diament and Goldsmith	A Multiple-choice examination in physiology. B.J.Med. Ed. Sept. 1969.
Bloom et al	Taxonomy of Educational Objectives. Vol. 1 1956.
Strawson	Introduction to Logical Theory 1952.
Wittgenstein	Philosophical Investigations 1953.