1

DOCUMENT RESUME

ED 183 589 TM 010 108

AUTHOR Tomko, Thomas N.; Ennis, Robert H.
TITLE Evaluation of Informal Logic Competence. Rational Thinking Reports Number 3.
INSTITUTION Illinois Univ., Urbana. Bureau of Educational Research.
PUB DATE 79
NOTE 55p.

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Cognitive Tests; Criterion Referenced Tests; *Critical Thinking; Evaluation Methods; *Logical Thinking; Norm Referenced Tests; State of the Art Reviews; *Student Testing; *Test Construction; Test Reliability; *Test Reviews; *Test Selection; Test Validity
IDENTIFIERS Cornell Critical Thinking Test; Watson Glaser Critical Thinking Appraisal

ABSTRACT

A discussion of evaluating informal logic competence centers on the identification of currently available tests and on a description of test theory, to aid in test selection. The following concepts in test theory are discussed: norm-referenced and criterion-referenced tests, true scores, reliability, validity, test selection and evaluation, test construction, and test format. The use of tests to assess student performance and in various experimental designs is also explained, as well as responsive evaluation, semi-structured evaluation, surveys and questionnaires, and longitudinal follow-up studies. (MH)

EVALUATION OF INFORMAL LOGIC COMPETENCE

BY

THOMAS N. TOMKO
AND
ROBERT H. ENNIS

RATIONAL THINKING REPORTS
NUMBER 3

General Editor of Rational Thinking Reports
Robert H. Ennis

Published by

The Illinois Rational Thinking Project

Bureau of Educational Research
University of Illinois at Urbana-Champaign
Urbana, Illinois
1979

Evaluation of Informal Logic Competence[1]

Thomas N. Tomko and Robert H. Ennis
Bureau of Educational Research
University of Illinois, Urbana-Champaign

One of the distinguishing features of the developing informal logic movement is its concern with pedagogy. Informal logic teachers want to know what skills and concepts are important for their students to know, how these skills and concepts can best be taught, and how one can determine if they have been successfully taught. Certainly the answers to the latter two questions will require some empirical research. In this paper we will examine tests and evaluation procedures in the hope that we can shed some light on these two questions. Our intended audience is people who have the above concerns, but who are unfamiliar with available tests and/or with the field of testing.

Tests

One naturally turns to tests as a starting point for the evaluation of teaching. Although not all people agree on the value of tests, they are a very practical and widely-used evaluation tool.

Currently available tests

The Watson-Glaser Critical Thinking Appraisal. This test is perhaps the most widely-used instrument in the area of logic and critical thinking. It has two forms, Ym and Zm (revised in 1964), each 100 items long and divided into five subtests:

Inference - ability to discriminate among degrees of truth, falsity, or probability of inference drawn from given facts or data

Recognition of Assumptions - ability to recognize unstated assumptions in given assertions or propositions

Deduction - ability to reason from given premises, recognition of logical implication

Interpretation - ability to weigh evidence and to discriminate among degrees of probable inference

[1]We would like to thank Robert Linn, Stephen Norris, Denis Phillips and Bruce Stewart for their comments on and criticism of a draft of this paper.

Evaluation of Arguments - ability to discriminate between strong and weak, important and irrelevant arguments

The type of answer to be selected varies with the sub-tests. For example in Test 1 (Inference) students must decide, after reading a paragraph, whether some proposed inferences are true, probably true, probably false, false, or cannot be judged due to insufficient data. In Test 2 (Recognition of Assumptions), the examinee must decide, given a statement, whether or not another statement is "presupposed or taken for granted". For example, given the statement "Let us immediately build superior armed force and thus keep peace and prosperity," one must decide whether the proposed assumption "The building of superior armed force guarantees the maintenance of peace and prosperity" is "necessarily" made or not made. In Test 3 (Deduction) and Test 4 (Interpretation), one must decide whether a conclusion does or does not follow. In Test 5 (Evaluation of Arguments), one must decide whether short, two-or three-sentence arguments are strong or weak. The questions in each test are preceded by sample items and an explanation of what the student is to do.

The test does not cover some aspects of critical thinking which one might wish to cover. For example, no semantical skills or concepts are covered, so there are no questions dealing with definition or ambiguity. There are no questions dealing with the reliability of observation statements or statements made by authorities.

To the extent that the Watson-Glaser test measures the ability to deal with induction, it has problems that also trouble the Cornell Critical Thinking Tests (CCTT), discussed below. In order to answer some of the items correctly, examinees must have certain background knowledge. If they do not, they may answer incorrectly even though the way they reached the

answer might be judged to be acceptable. This problem appears in just about any induction test. We have found that it is very difficult to construct test items on inductive reasoning that have one answer that is clearly the best. One can almost always defend alternative answers by making certain reasonable assumptions that were not foreseen by the test maker.

Another problem with the Watson-Glaser test is its concept of an assumption. The directions to the section, "Recognition of Assumptions", say, "If you think the assumption is not necessarily taken for granted in the statement, make a heavy line under 'ASSUMPTION NOT MADE' on the Answer Sheet." (p. 4) We can probably neglect the problem of referring to the proposed assumption as an "assumption", but there is a more serious problem. The wording encourages the test taker to look for something that is necessarily taken for granted. Possibly presuppositions (of the type discussed by Strawson, 1952) are "necessarily" taken for granted, but premise-type assumptions are not. There is always another and different premise or premises to fill a real gap in an argument. So no premise-type assumption is necessarily taken for granted (Ennis, 1961, elaborates this point).

A third problem is that one's answers to the items in Test 5, Evaluation of Arguments, often depend on one's politics and values. It does not seem fair to mark a person wrong because of the person's politics and values. For example, Items 89 and 91 are keyed weak arguments on the issue of whether the United States government should try to inform the public of sought-after results of its scientific research programs in the areas of "new weapons, equipment, devices, etc.". These items are:

> 89. No; some people become critical of the government when widely publicized projects turn out unsuccessfully.
>
> 91. Yes; the projects are supported by taxes and the general public would like to know how their money is to be spent.

We agree with the key on Item 89 and disagree on Item 91, but we can see how people with a value position different from ours might well answer just the opposite. We are not here urging a relativistic ethics, but do not think that a person's critical-thinking score should depend on that person's politics and values in such cases.

A fourth problem for some is the orientation of the test to the United States, exemplified in the above-mentioned issue in Test 5. Students from other English-speaking countries might be penalized for being less familiar with United States government structure, policies, and problems.

The Cornell Critical Thinking Tests. These are actually two different tests, Level X and Level Z (Ennis & Millman, 1971a, b, c). They are not parallel forms of the same test, although they are both general measures of critical thinking ability. Level X is meant to be used for testing students roughly from junior high school through first year college age. Level Z is intended for testing examinees of college age on up, although high ability secondary students could also take the test.

The rationale for the test is based upon a concept of critical thinking set forth in Ennis (1962). According to the test manual,

> A critical thinker is characterized by proficiency in judging whether:
>
> 1. A statement follows from the premises.
> 2. Something is an assumption.
> 3. An observation statement is reliable.
> 4. An alleged authority is reliable.
> 5. A simple generalization is warranted.
> 6. A hypothesis is warranted.
> 7. A theory is warranted.
> 8. An argument depends upon an ambiguity.
> 9. A statement is overvague or overspecific.
> 10. A reason is relevant. (Ennis & Millman, 1971a)

Not all of these aspects of critical thinking are covered by each test. Level X does not cover aspects 7, 8, and 9, while Level Z does not cover aspect 7.

In the Level X test, examinees read a story about space explorers on a distant planet and are asked questions as the story unfolds. For example, at one point the explorers are said to be watching a group of beings who are standing around a campfire. In one type of question, the examinees must decide whether the first of the following underlined statements is more reliable than the second, the second more reliable than the first, or whether neither is more reliable than the other.

A. The mechanic, looking through his field glasses, says, "<u>They are tan-skinned creatures with furry spots.</u>"

B. The anthropologist, looking through his field glasses, says, "<u>They don't have furry spots. They are wearing the skins of animals.</u>"

C. Equally reliable or unreliable.

The Level Z test is somewhat more difficult and the questions and directions are more complex. For example, after an experiment and its results are described, examinees are asked whether certain additional information, if true, would make the results a) more certain, b) less certain, or c) neither. There is also a section on semantic skills and concepts not covered on the Level X test.

Like the Watson-Glaser test, the Level X test has its problems with induction. It tries to avoid some of these problems by asking whether proposed evidence counts for, against, or is neutral with respect to a certain hypothesis; that is, it only asks in which direction the evidence points. But as in the case of the Watson-Glaser test, one can, by making reasonable assumptions, dispute the keyed answers. We feel that this is

the major problem that people writing tests of inductive-reasoning ability must overcome.

Some of the subsections of the Level Z test seem a little short! For example, the section on judging the reliability of observations and authorities is only 4 items long, as compared with 21 items on the Level X test.

The Watson-Glaser Critical Thinking Appraisal, the Cornell Critical Thinking Test, Level X and Level Z are, in our opinion, the only general tests of critical thinking currently available. But there are two other available tests that claim to measure or appear to attempt to measure general critical thinking ability, and there is a set of "indexes" that might jointly be construed as a critical thinking test.

One test is The Uncritical Inference Test by William V. Haney (1975). It consists of three stories, each followed by a series of statements (76 items altogether) which the examinee must decide are either "definitely true", "definitely false", or "?" (questionable) based on the information given in the story. Due to the large number of answers keyed "?", a hardened skeptic or "pathological doubter" would receive a high score on this test. (This was a problem with an earlier version of the Watson-Glaser test, also. See Ennis, 1958). The test appears to be essentially a test of whether a person is willing to infer at all beyond what is very explicitly stated, and might better be called a critical uninference test.

The other is entitled Logical Reasoning, developed by J. P. Guilford and A. F. Hertzka (1955), designed to assess what Guilford calls the factor of evaluation of semantic implications. Although he claims that this factor is known as critical thinking ability, the test itself consists of forty items, each of which presents two premises of a syllogism and asks the examinee

to pick the correct conclusion from four possible alternatives. This seems to be a test of only one aspect of critical thinking, viz., class reasoning.

Then there are several aspect-specific tests of critical thinking ability. The Cornell Conditional Reasoning Test (Ennis, Gardiner, Guzzetta, Morrow, Paulus, & Ringel, 1964) and the Cornell Class Reasoning Test (Ennis, Gardiner, Morrow, Paulus, & Ringel, 1964) are designed to test the abilities named in their titles. The Evaluation Aptitude Test, by D. E. Sell (1952), consists of 36 syllogisms. The test is divided into two parts, one containing neutral items and the other containing emotionally-loaded items. The test is meant to be used, in part, to measure the degree to which emotional bias influences deductive-reasoning ability.

A set of aspect-specific tests has been constructed by the Instructional Objectives Exchange (IOX), formerly associated with the Center for the Study of Evaluation at UCLA. It is an organization that serves as a clearinghouse for instructional objectives and tests designed to measure whether or not students have attained those objectives. Each test (or "index", as they put it) is meant to measure the attainment of one or more stated objectives.

In Judgement: Deductive Logic and Assumption Finding (Instructional Objectives Exchange, 1971), seven objectives from the area of informal logic are presented in conjunction with five indexes designed to measure the attainment of those objectives. The Conditional Reasoning Index and the Class Reasoning Index are meant to measure two objectives each. The work of Ennis and Paulus (1965) is cited as a basis for the items included, but one should note the sense in which the IOX authors use the term 'valid'. The directions ask examinees whether a proposed conclusion is valid or invalid, given one or more premises, rather than whether an argument, or line of reasoning, is valid.

Two other objectives and their corresponding indexes deal with assumption-finding. However, the term 'assumption' has several senses and this could cause some problems if one is not alert. Objective 5 in IOX (1971, p. 4) states the following:

> Given a series of statements, each of which is followed by several proposed assumptions, the students will determine whether, within each question set, each of the assumptions listed is necessary to the particular statement.

Assumption Recognition Index I, is the test associated with this objective. Objective 6 on the other hand, involves judging whether statements are necessary to arguments, a skill sometimes called "gap-filling". Assumption Recognition Index II is the test associated with objective 6.

After being presented with a statement in the first test and an argument in the second test, the examinee is asked to decide whether a person who offers an argument or statement must accept each of a series of additional statements in order to be "reasonable and consistent". In Assumption Recognition Index I, the keyed answers appear to be either (Strawsonian) presuppositions or (Gricean) implicatures, and in Assumption Recognition Index II, the keyed answers are either gap-fillers or implicatures. While one might argue that presuppositions are, in some sense necessary for certain statements and that certain gap-fillers are needed (though not logically necessary) for some arguments (given a context), it is not clear that implicatures are necessary for either statements or arguments. In addition, using the label "assumption recognition" for tests which involve identifying presuppositions, implicatures, and gap-fillers may be misleading to the test user searching for a test to cover a homogeneous notion of assumption recognition.

The final objective and its accompanying test, Recognizing Reliable Observations, is also based on the work of Ennis (1962). The items in this

test seem clear and well-written, but given the number of criteria listed by Ennis for judging the reliability of observation statements, there is some question as to whether the test is long enough (10 items). This test and the other IOX tests could be useful aids for the informal logic teacher, but, as is the case with any test that one has not constructed oneself, they should be examined carefully before they are used. There is no discussion of cut-off scores, nor any argument offered for the representativeness of the items.

The tests mentioned above are the ones we have found that are still in print.* They are contained in a critical thinking test file we are developing at the Illinois Rational Thinking Project, for which we hope to obtain copies of all existing general and aspect-specific critical thinking tests. One Project member, Bruce Stewart, has written a collection of reviews of tests in the areas of critical thinking and informal logic (Stewart, 1979), containing reviews of about thirty tests.

Another source for information about tests of all sorts is Buros' Mental Measurements Yearbook (1972). Buros lists tests in many areas and also includes reviews of some. The 1972 edition is now somewhat out of date, but a new edition should appear soon.

*Perhaps Sell's Evaluation Aptitude Test is out of print now. It is not listed in the most recent Psychometric Affiliates catalogue.

Concepts of Test Theory

Although the selection of tests of interest to the teacher of informal logic is unfortunately somewhat limited, it is useful to know what criteria are generally used to choose a test. An understanding of these criteria and an acquaintance with their theoretical background can help one get a better overall picture of the process of evaluation of teaching and research in informal logic.

Norm-referenced and criterion-referenced testing. A test that is intended to measure the absolute standing of individuals with respect to some standard of performance (often mastery) is generally called by test theorists a criterion-referenced or content-referenced test. ('Criterion-referenced' is a term introduced by Glaser (Glaser & Klaus, 1962).) Its inclusion of the word 'criterion' is somewhat unfortunate, since 'criterion' has another use in test theory that could cause some confusion. (See the discussion of predictive validity below.)

Someone might alternatively be interested, not in assessing degree of achievement, but rather in determining differences among groups of students and individual students. A researcher making compatisons would use a test that assesses the relative standing of individuals with respect to the possession of a trait or traits. This type of test is generally called by test theorists a norm-referenced test. A person's score on such a test is an indication of how well he or she performed as compared with other individuals who took the test. Someone who scores at the 90th percentile of a norm-referenced test has done as well or better than 90% of the people with whom he or she is being compared. Such scores, however, do not indicate whether the level of mastery was low or high. So, for example, someone who

scores at the 90th percentile on a test of reading ability may not be a very good reader. He or she is just better than 90% of the group with whom he or she is being compared (the norm group).

Since it appears that the same test can be used both as a criterion-referenced test and a norm-referenced test, we propose to change the labels slightly and talk of criterion-referenced testing and norm-referenced testing. This labeling explicitly recognizes the dependence of the distinction upon purpose and interpretation in the given situation. This relabeling relieves us of the burden of classifying every test as one or the other type, a task we have found in practice to be impossible.

At the present time, the bulk of published tests are developed and defended for the purpose of norm-referenced testing. Examples of such tests include IQ tests and college entrance exams, such as the Scholastic Aptitude Test (SAT) and the Graduate Records Exam (GRE). The Watson-Glaser and Cornell Critical Thinking Tests might be usable for either purpose, depending on whether they are judged by a test user to be adequate in a particular situation. The IOX tests are designed to be used for criterion-referenced testing. Competency testing, which is becoming popular in elementary and secondary schools, is a type of criterion-referenced testing. (For a discussion of problems with competency testing, in addition to the general problems with criterion-referenced testing that we will mention, see Smith, 1975.)

The fact that a test was designed for norm-referenced testing does not preclude its use for criterion-referenced testing (or vice-versa). However, some knowledge of the theories or models behind a test is helpful in deciding whether a particular test is appropriate for a use one has in mind.

Most of the testing terminology that we shall introduce was originally used in the context of classical mental test theory, which was developed to cover norm-referenced testing. It should be noted, however, that norms are not necessary for the employment of classical test theory (see Lord & Novick, 1968, p. 34, for the assumptions of classical test theory). The key concept in classical test theory is variance, not norms. In fact, many of the terms of classical test theory can be used when discussing criterion-referenced testing, but it is not clear that all of the terms make sense when so employed.

True scores. Although there are now several types of mental-test score models, classical test theory is what is known as a true-score model. On this model, an individual's observed score on a test, X, consists of a true score, T, plus an error score E. A partly philosophical problem, the nature of true scores, is discussed by test specialists, Frederick Lord and Melvin Novick (1968, pp. 39-44). Lord and Novick also discuss the basic assumptions of classical and other test theories.

Reliability. The concepts of reliability and validity are very prominent in the literature on testing. By definition a test is reliable to the extent that it produces consistent results from one application to the next; and, roughly speaking, a test is valid to the extent that it measures (or correctly appraises) what it is supposed to measure (or correctly appraise). These are very rough definitions, but they do give one an intuitive handle on the concepts. Both concepts are problematic in application.

In contrast to discussion of all but two types of validity, discussions of test reliability generally have at least the appearance of precision, as a result of the complex statistical techniques that are employed to investigate the reliability of tests. Despite their complexity, however, these techniques

are much easier to deal with than the controversial procedures of test validation. These facts may explain why there is an inordinate amount of emphasis placed on establishing test reliability as opposed to showing the validity of a test. In choosing tests, one is likely to encounter data on test reliability rather frequently.

As defined above, a test is reliable to the extent that it produces consistent results from one application to the next. This is similar to what one would expect in a theory of measurement in the physical sciences. A ruler is reliable to the extent that it produces a consistent set of data from one application to the next. To determine whether a ruler was reliable, one might repeatedly measure the same thing. A reliable instrument would produce very close readings on the repeated measurements. For some types of tests covered by educational test theory, such as physical or motor skills tests, this notion of reliability will suffice. But for most educational tests, reliability cannot be determined in terms of direct comparison of repeated measurements of the same individual using one instrument. This is partly because human beings are often changed by the measurement process itself. The act of taking a test can affect the trait being measured by the test. Consequently, one would not expect consistent scores on repeated measures. In fact, one might find artificially consistent scores, since examinees sometimes remember their original responses when taking the retest.

In an attempt to surmount this problem, test theorists employ the notion of a parallel test form. Parallel test forms are defined as tests that produce parallel measurements. Two measurements are said to be parallel measurements if each individual's true score on the two measurements is the

same and if the variance of the error scores for the two measures are equal (that is, the error scores for the two measures are spread out to the same extent). What this notion does for test theorists can be seen in the following quote from a standard text in this area, Statistical Theories of Mental Test Scores by Lord and Novick (1968, p. 43): "Thus parallel measurements measure exactly the same thing in the same scale, and, in a sense, measure it equally well for all persons."

The reason that parallel measures are used in discussions of reliability is that one common method of defining reliability involves the unobservable quantity, an individual's true score; that is, reliability is defined as the squared correlation between observed score (X) and true score (T): ρ^2_{XT}. It deductively follows from the assumptions of the theory, however, that $\rho^2_{XT} = \rho_{XX'}$, where X and X' are parallel measures and are potentially observable. So the concept of parallel forms is helpful in developing a usable theory of reliability.

It is very often judged too difficult, too time-consuming, or too expensive to develop a parallel form for a test. When this is the case, there is a third, widely-used approach to estimating the reliability of a test: estimating the internal consistency of the test. These estimates use only a single test form to estimate reliability. One such procedure is the split-half method. In this method one first divides the test into two halves that are assumed to be parallel. (There are, of course, many ways to split a test. How the splitting is done depends upon the nature of the test.) The scores on the two split-halves are then correlated. This does not produce an estimate of the reliability of the original test, but of a test only half

as long. To estimate the reliability of the original test, one uses a formula, the "Spearman-Brown" formula, that estimates the reliability of a test that is longer than a given test.

In addition to split-half internal consistency, there is another common single-form approach to estimating the reliability of a test. This approach we shall call the multiple internal-consistency approach, since it relies on the extent to which all of the items intercorrelate with each other. The Kuder-Richardson formulas (KR-20 and KR-21) are commonly used ways of estimating multiple internal consistency.

There are other methods of estimating reliability by looking at the internal consistency of a test, and one can find a discussion of these approaches, as well as a cogent, but somewhat technical, discussion of the concept of reliability by Julian C. Stanley in a book chapter entitled, "Reliability" (1971).

There is a significant problem in the use of multiple internal-consistency reliability estimates. Informal logic and critical thinking are probably heterogeneous notions, but multiple internal-consistency reliability gives higher ratings to homogeneous tests. Hence in building tests to produce high multiple internal-consistency reliability there is the tendency to eliminate items that do not correlate highly with the rest -- even though such items may be very good indicators of some feature that is not highly correlated with the other features of these heterogeneous notions. Such items tend to be eliminated by test makers interested in making the reported numbers look good, and the reason that such items are eliminated is simply that they are non-conformist.

This problem of using multiple internal-consistency as a substitute for the original notion of reliability (consistency of repeated applications) is not peculiar to the domain of testing in informal logic and critical thinking. It permeates most mental testing by highly respected organizations. Informal logicians are especially suited to guard against the resulting invitation to equivocation in defense of a test. We urge their sharing of this insight with others less sensitive to equivocal arguments.

Validity. It is not enough for a test to given consistent scores; it also must measure what we desire to measure. The process of determining whether a test measures what it is designed to measure is called test validation. When sufficient evidence has been accumulated to support the claim that the test measures a certain variable, the test is said to be a valid measure of that variable. Actually, this way of speaking is slightly misleading, although it is often encountered. Cronbach (1971, p. 447) urged that it is not the test instrument per se that is valid, but the interpretation of data arising from a particular use of the test. A single test can be used in many ways (e.g., research, placement, job-screening, grading, etc.). Some interpretations for a particular use may be valid; other interpretations of data from the same test used for different purposes may not be valid.

Although Cronbach's point is an important one, most discussions of tests are still carried on with references to the "validity of a test." One should understand such locutions as essentially incomplete expressions. A test's validity must be understood as its ability to measure or be a sign of one or more specified things when it is given under particular conditions. These additional qualifications must be kept in mind if one encounters talk of the validity of a test instead of talk of the validity of interpretation of test scores.

Despite the appearance of precision given by the statistical superstructure of test theory, many of the central concepts in this field are not at all precise or well-clarified. Validity is one such concept. There are actually several somewhat loosely-related concepts which come under the heading "validity". We shall briefly characterize five of the concepts that one is likely to encounter in discussions of testing.

A test is said to have face validity if it appears to be a valid test. According to the American Psychological Association's Standards for Educational and Psychological Tests (1974), face validity is the "mere appearance of validity" (p. 26). Most test experts feel that face validity is illegitimate, but we do not see how to get along without it. It appears to be an essential element of content validity.

To explain their notion of content validity, test theorists introduce the concept, universe of behaviors. Such a universe consists of a set (possibly infinite) of behaviors that a student should be able to exhibit if he or she has grasped certain concepts or mastered certain skills. This concept, universe of behaviors, is ripe for philosophical inquiry. Although we shall use it in presenting established theory because it is part of established theory, we have many reservations about its use.

We often design tests to determine whether our students have learned what we intended to teach them. For example, after a unit on propositional logic, we want to be able to determine whether our students learned something about that topic. We cannot, of course, test for all possible "behaviors" we would expect a successful student to be able to exhibit. We are expected to try to get a representative sample of the universe of behaviors for which

we want to test. How one can (without leaning heavily on face validity judgments) identify a representative sample of a critical-thinking universe of behaviors is, unfortunately, unclear. However, a test is said to have content validity to the extent that we actually did choose a representative sample. It does seem that a test on propositional logic that only asked questions of the following form would not have much content validity: Is the following argument valid?

If r, then s.

r.

Therefore s.

Assuming that the only difference among the items is the use of different single letters in the place of 'r' and 's', the items do not appear to call for a representative sample of propositional logic behaviors, in whatever way we choose to interpret the word "behaviors". But that judgment appears to be a face validity judgment. If not, then we would have had to have a way of describing and identifying the things in a (infinite) domain of propositional logic behaviors and of drawing a random or systematic sample from that domain. This describing, identifying, and sampling makes no sense to us, but we invite other philosophers to work on the problem.

The resolution of this difficulty is important since the clarification of the concepts of face validity and content validity is essential for the further development of theories applicable to criterion-referenced testing. Such testing is of paramount interest to those who wish to evaluate the extent to which students have mastered the content of informal logic courses.

Two further types of validity, distinct from the previous two but related to each other, are predictive validity and concurrent validity. In many

cases, a tester is interested in estimating the value of a certain variable from the score on a test. For example, college admissions officers would like to estimate a candidate's freshman grade point average, given the candidate's score on an entrance exam, such as the SAT or ACT. The variable to be estimated is called the criterion. (This kind of criterion should not be confused with that in criterion-referenced testing. Predictive and concurrent validity are not usually concerns in criterion-referenced testing.) A test has predictive validity if knowing a subject's score on the test enables us accurately to predict the value of the criterion. The difference between predictive and concurrent validity lies in the temporal relation between the test score and the criterion. One speaks of concurrent validity when one is interested in a subject's standing on the criterion at the time of test administration, while predictive validity is the concern when one is interested in the subject's standing on the criterion at some future time.

Concern with predictive validity at one time dominated test theory, although at the present time more attention is being given to construct validity than was given to it in the past. One is investigating the construct validity of a test when one attempts to confirm or disconfirm that a test measures some hypothetical, unobservable psychological construct. Intelligence, anxiety, and critical thinking ability are examples of such constructs. The claim that a test measures intelligence, we feel, is a claim about the construct validity of a test, as is the claim that a test measures critical thinking ability. (Behaviorists, who reject the idea of construct validity, would of course not agree.)

The investigation of construct validity can be viewed as the process of placing the specified construct in the context of some larger theory,

and ascertaining the acceptability of the theory, the role the construct plays in the theory, and the relationship between the test and the construct.

> Construct validation is the process of marshalling evidence in the form of theoretically relevant empirical relations to support the inference that an observed response consistency has a particular meaning. (Messick, 1975, p. 55)

Construct validity is viewed by some as a broad concept that encompasses or subsumes all other types of validity.

The model employed in construct validation is the neo-positivist theory of confirmation as set forth by Carl Hempel (1965, 1966). Statements about the construct in question are located in a hypothetico-deductive system. Predictions are deduced from the system and either confirmed or disconfirmed. The construct validity of the approach to interpretation of test scores is supported to the extent that the predictions are confirmed and to the extent that the predictions depend upon the relationships among the test, the construct, and the other elements of the system (which are also constructs). Philosophers are well-acquainted with the extensive and powerful criticisms of this model, but such criticism have as yet had little effect on the use of hypothetico-deductive models in test theory. (This is not to say that test theorists are completely unaware of the problems involved, cf. Cronbach, 1971.) A standard philsophical problem connected with construct validity is the nature of the constructs being investigated. For example, what (if anything) is being referred to by the term "critical thinking ability"?

Judging tests. How should one use the concepts discussed above when judging an available test? That depends to a great extent on the purpose for which one is using the test. Some remarks about general cases, however, can provide some guidance. One must take care when using the information

provided about tests, since, as we have to some extent indicated, there are problems involved with theory behind tests and their interpretation.

One very general problem involved in making judgments for criterion-referenced testing is the source of the vocabulary which is used to talk about tests. Classical test theory was developed to cover norm-referenced testing. Some of the terms which are appropriate to use when discussing norm-referenced testing are not clearly applicable to criterion-referenced testing. There is as yet no theory of and vocabulary for criterion-referenced testing comparable to the theory and vocabulary which have been developed for norm-referenced testing, although progress has been made in this area during the last decade. (See Hambleton, Swaminathan, Algina & Coulson, 1978.) Nevertheless, some test theory concepts seem applicable to both types of testing.

Reliability as a basis for judgment. High reliability seems to be desirable for any test although, as indicated earlier, the commonly-used internal-consistency formulas for estimating reliability are misleading for nonhomogeneous tests. There are no firm requirements for reliability coefficients, although some have been suggested. A widely-quoted set of minimums was set forth by Kelley (1927):

a)	To evaluate level of group accomplishment	.50
b)	To evaluate differences in level of group accomplishments in two or more performances	.90
c)	To evaluate level of individual accomplishment	.94
d)	To evaluate differences in level of individual accomplishment in two or more performances	.98

These figures of course depend on the assumptions Kelley made about required fineness of discrimination and the acceptable chances of going wrong. They also partly explain why many professional testing people are so devoted to obtaining high reliabilities: Here we have some "objective" standards, and

there are test development procedures that generally will enable one to meet these standards--at a cost. The cost might be 1) excessive testing time, 2) excessive demands on the time of experts, 3) triviality of items, and 4) neglect of important features of the trait(s) for which one is testing. (Remember the pressure for homogeneity of items resulting from the use of internal consistency formulas for reliability estimation.) To the extent that we accept these last two costs, a frequent occurrence, we get a reliable, but invalid test.

Heavy reliance on reliability is also related to the fact that early test developers were often interested in predicting the standing of a subject on some criterion. They were concerned with predictive validity. According to Lord and Novick, reliability can be viewed as predictive validity with respect to a parallel test (1968, p. 63). Consequently, reliability was seen as part of the only concept of validity thought to be important, criterion-related validity.

One way to increase multiple internal-consistency reliability is to secure item homogeneity. Another, according to classical test theory, is simply to increase the number of items on the test. To illustrate this phenomenon, consider the Cornell Critical Thinking Test, Level X and the Cornell Critical Thinking Test, Level Z. Depending on the group from which the data was collected, the estimated reliabilities of the tests range from .77 to .87 for Level X and from .55 to .77 for Level Z. However, Level Z is only 52 items long, as compared with 71 items on Level X. Using the Spearman-Brown formula, one can show that if Level Z were as long as Level X, its reliability estimates would range from .62 to .82, which is closer to the range for Level X.

Most currently-used procedures for estimating reliability, even including split-half methods, do not capture the full-blooded notion of reliability. They are based on one test administration. Consequently instability of measurement over repeated administrations is not taken into account. (For a more detailed explanation of this problem, see Cureton, 1965. Also, see section F of the American Psychological Association's Standards for Educational and Psychological Tests (1974).)

The concept of reliability must be cautiously applied to criterion-referenced testing. Some techniques which can be used to increase the reliability for norm-referenced testing are not appropriate for criterion-referenced testing. For example, when revising a test, the reliability can be increased if one retains items with a "difficulty index" of about .5, meaning that the proportion of examinees obtaining the correct answer is .5. (This index is misleadingly named. It might better be called the "ease" index", as suggested by Ahmann & Glock, 1958.) If instruction has been effective, the difficulty (read "ease") index should be high for items on a test for criterion-referenced testing, meaning that a high proportion of students should answer the items correctly. A test maker who aims for items with a .5 difficulty index will then be forced to construct overly difficult, recondite, or nit-picking questions.

In summary, there are a number of traps facing someone pursuing high reliabilities, and in accepting the judgments of others who pursue high reliabilities.

Validity as a basis for judgment. In almost all cases of interest to the teacher or researcher in informal logic, one must also ask whether a test is valid. As a first step in judging the validity of a test, one should

examine the description of the test that should be in the manual to see whether the test comes close to what one seeks. Then, if it appears worthwhile to go on, one should scrutinize the items very carefully. The best way to do this is to take the test under the prescribed conditions and check one's answers against the key, seeking for explanation and resolution of any discrepancies. After going through this process, you will have a fairly good idea about the extent to which the test does what you want it to do. A judgment based on such an inspection would be a judgment about the so-called "face validity" of a test. Going through these steps makes good sense, even though face validity is a disreputable notion in the eyes of many test theorists, making this low regard somewhat puzzling.

Judging a test for its content validity, as defined above, requires that one adopt and employ the concept universe of behaviors. As indicated earlier we shall provisionally do so for the purposes of applying this approach.

Judgments about the content validity of a test should be aided by the examination of the test rationale that should appear in the test manual. This rationale should somehow help one identify all the members of the universe of behaviors, so that one can then decide whether the test items call for a representative sample from it. How one actually does all this we do not know. In one sort of actual practice it appears that content validity is established by making the topics in the rationale quite specific and, if possible, by transforming them into types of behavior to be exhibited in types of situations (rather than into specific items of behavior*). This list of types of behaviors is called a table of specifications. Then face validity judgments are made (though they are not called face validity judgments) about the

*This distinction is between behaviors' being dispositions and their being performances.

item-produced behaviors' representativeness of the types of behavior in the table of specifications.

A second procedure for establishing content validity in actual practice is to gather a large number of items that an expert judges (another face validity judgment) to call for behaviors that are representative of types of behavior desired. Then a random sample of some sort (or a systematic sample) is drawn from the item pool, and the test consists of this sample.

Note that both of these procedures for establishing content validity do in fact lean heavily on face validity judgments. Content validity, in the only ways we can conceive of its pursuit, consists of organized systematic ways of utilizing the face validity judgments of experts. Both of the content-validity procedures we have outlined can be followed by someone building a test of logical competence, and can be evaluated for their care and quality by consumers of such tests.

These processes of judging pure face validity and content validity are applicable to any critical thinking test, whether for norm-referenced testing or criterion-reference testing, and whether the test was originally constructed for norm-referenced or criterion-referenced purposes.

A problem that has still not received much attention in the literature is the problem of making judgments about desirable levels of performance for criterion-referenced testing. What level of performance should be considered evidence of mastery? This is a difficult question to which developing theory does not yet have an answer, even though the test user often seeks an answer to this question. (See Popham, 1971 for a sympathetic discussion of problems in the theory of criterion-referenced testing.)

Establishing the construct validity of tests is a difficult task, in part because <u>construct validity</u> in itself is not a crystal-clear notion.

Much more attention has been given to construct validity in the past several years than was given to it in the early days of the development of test theory. But problems still remain.

As described above, the process of making a case for the construct validity of a test consists of showing how the construct fits into some larger theory. One way to do this is to show how scores on the test in question are related to other variables. So, for example, one would expect that critical thinking ability, since it involves judgments about statements, would be moderately related to reading ability. Many test manuals offer lists of correlations of the test in question with other variables. But such a list by itself does not establish the construct validity of a test. One must show how the correlations would be expected to follow from a theory in which the construct in question is embedded.

One important place to look for evidence of construct validity is in the relation between a test and other closely related measures. For example, one would expect a high correlation between tests which claim to measure critical thinking ability. Such "convergence of measures" gives some support to the construct validity of all of the tests involved. Lack of agreement could mean several things: a poorly constructed test, differing conceptions of critical thinking, unshared prerequisite familiarity with the subject matter, etc.

One might also expect the constructs measured by a particular text to be unrelated to certain other constructs, that is, one should be able to discriminate between unrelated constructs. Tests measuring unrelated constructs should be weakly correlated or uncorrelated (see Campbell & Fiske, 1959).

Construct validity arguments for existing critical thinking tests are either weak (The Cornell Tests and the Watson-Glaser test) or nonexistent (the others). This is a comment about the arguments for construct validity, not about the construct validity of the tests.

There are certain positions of which one should be aware when reading discussions about construct validity. One may encounter those who demand a reductionist operational definition of each construct. The strict operationalist does not view such a definition as providing a method of measuring the construct in question, but thinks that each test defines a different construct. There are many criticisms of this view (for example, Ennis, 1964), and even neo-positivists such as Hempel (1961) regard such a position as too rigid, but one frequently encounters this position in discussions of educational testing. Holders of this position are opponents of the use of construct validity in test appraisal (for an example, see Bechtoldt, 1959).

Another position that one occasionally encounters holds that high correlation implies conceptual identity. That is, if two tests correlate highly then they measure the same thing. Cronbach proposes a counterexample to this position (1969): Comprehension of physical laws will correlate highly with scientific reasoning ability, but this does not mean they are identical. It may simply be the case that, at present, the best curricula do a good job on both and the worst do a poor job on both.

Construct validity questions are usually associated with norm-referenced testing. Some experts in testing feel that one need not consider construct validity when assessing tests for criterion-referenced testing. However, there has recently been criticism of this position. The literature in this area is

just now developing and little can actually be reported at present, but it is an area that merits watching and participation by philosophers as it develops.

Sometimes predictive and concurrent validity will be of use to informal logicians, and when they are, the goal is high correlations between the test and the criterion. Generally correlations between scholastic aptitude tests and levels of later subject matter achievement run about .5 (this is predictive validity); correlations of tests with other tests that are testing for the same thing go up to .8 when the tests are fairly similar (this is concurrent validity when the other tests are administered at roughly the same time). These numbers might serve as rough guides for what one can expect. Statistical significance of correlations is generally of little interest for predictive and concurrent validity, since that standard is too easy to satisfy with a sample of any reasonable size.

The major problem in using predictive and concurrent validity in evaluating informal logic tests is that of finding a criterion that can justifiably be assumed to be better than the test in question.

The suggestions given above for assessing tests are not meant to be exhaustive. Many other considerations enter into the choice of a test (e.g., time limits, cost, reading level).

Constructing Tests

After examining available informal logic tests, one might conclude that none are appropriate for the purposes at hand. At this point a natural move would be to consider the task of constructing one's own test.

Constructing a good multiple-choice test is no easy task. It cannot be accomplished at one sitting, since, ideally, test construction involves several distinct time-consuming steps. One might object that the time

and effort involved would not be justified if one just wants to make up a mid-term exam. While we might not ordinarily undertake a grand project in such a case, nevertheless, attention to the procedures outlined below can improve the quality of many tests.

As one would expect, the procedures for constructing instruments for norm-referenced and criterion-referenced testing are somewhat different, although one instrument might serve both purposes. Each of the following list of procedures is an edited and abridged version of a presentation in Sax, 1974. If you follow these procedures, it is essential to ask frequently whether what you are doing makes sense. Mechanical rule following is dangerous, but an easy trap to fall into. If test specialists are employed, the informal logician must monitor the process closely.

Constructing tests for norm-referenced use. The following steps describe the procedures one would usually follow in constructing a test for norm-referenced use.

1. Test rationale and objectives are determined. This serves as a foundation for the writing of items. It also serves a part of the case for the face, content, and construct validity of the test. The content of the test is determined by the nature of the field in some tests and by the type of objective to be tested for in others. Subject matter experts should be involved.

2. Next, items are written for the test. More items are written than will be included in the final form of the test. Sometimes several preliminary versions are constructed. Although the multiple-choice format is most often used, other formats are possible (see "Test format" below).

3. The items are administered and the results are analyzed. It is desirable to give the proposed items to a fairly large and varied sample of the population for which the test is designed. For some widely used tests, such as the SAT, tens of thousands of examinees take the trial tests, but much smaller samples can be used. Most colleges and universities and some high schools now have computer facilities which give detailed item analyses for machine-scored tests.

Two standard results of an item analysis are the difficulty index (described earlier), and the discrimination index. The discrimination index is an attempt to indicate how well an item distinguishes between two groups of people otherwise identified. Often these groups are the top and bottom groups on the total score on the test; if so, then there is a danger of overemphasizing homogeneity and neglecting important aspects of informal logic, if any, that do not correlate highly with the ones that dominate the test. Seeking high discrimination indices based upon total score helps achieve high internal-consistency reliability estimates, a trap mentioned earlier. But in any case items with low discrimination indices should be carefully scrutinized. Often the cause is a problem in the wording of the item. Poorly worded distractors (supposedly incorrect alternative answers) that should not be scored "wrong" can often be detected by item analyses that indicate how groups selecting each distractor performed on the total test.

For norm-referenced testing, difficulty indices of .5 are often sought because that is a good way to spread people out on a continuum. Dangers of using this standard for criterion-referenced testing were mentioned earlier, and they apply to some extent to norm-referenced t ng as well. We might

succeed (by following this procedure) in spreading people out on a continuum, but the continuum might, as a consequence, be of little interest.

In any case, the results of item analyses, one should remember, apply at best to groups similar to the group that took the test. They might not apply at all to a different group. A danger here is to do an item analysis using a group that has received no instruction in informal logic, and then to use the test on a group that has had considerable instruction in informal logic. Opportunities for distortion abound. Reliability estimates should be computed, and validity evidence should be considered.

4. The final test form is constructed. Factors such as time limits for test taking influence the number of items included in a form.

5. The final form is administered to another large and varied group of examinees and normative data are generated for use in subsequent administrations. Widely used tests are continually being revised and norms are frequently updated.

At this point, one would have a test that is norm-referenced, but not necessarily a good test. If items are chosen with reliability in mind after step 3, the reliability of the test is likely to be high. Even if evidence regarding face, content, and concurrent validity is present, predictive and construct validity would still need to be determined. It should be apparent that the construction of a good test for norm-referenced purposes could take several years.

Constructing tests for criterion-referenced testing. In constructing such tests, one follows procedures similar to but not identical with norm-referenced procedures. One should also keep in mind that these tests do not have the backing theory which norm-referenced tests have.

1. A general test rationale is prepared

2. The universe of behaviors to be covered by the test is specified. These specifications indicate what a student who has mastered the universe should be able to do, although it is not clear whether the "behaviors" are to be dispositions or performances.

3. Test items are written which conform to the specifications of step 1. How to do this is not clear, because in most content areas, the nature of the relationship between the items and the universe of behaviors is not clear. Be that as it may, one next makes (ideally) a random selection from all such possible test items, but this is usually not possible since most universes subsume an infinite number of items. Instead, one might try to assure that the sample of items selected is representative by comparing the items with the universe specifications (a face validity judgment).

4. If one has to choose from among the items selected in step 2 (to adjust the test for proper time length, for example), those items that discriminate most clearly between those who have had instruction and those who have not are usually preferred, other things being equal.

5. Standards of competence are determined. There is controversy over whether this step can or should be taken. Although many criterion-referenced tests have cut-off scores, Glass (1978) argues that procedures used in determining cut-off points are indefensible.

6. The test is administered under conditions that conform to the universe specifications (i.e., if the universe of behaviors deals with written criticism of written arguments, an oral test would not be appropriate).

7. Student performance is assessed by comparing test results with the specified standards of competence. One checks to see whether the ratings of the students make sense.

As with norm-referenced testing, these procedures lend themselves to continuing test revision and improvement over time. Items that discriminate most clearly between those who have and have not mastered the material are retained and other items are scrutinized for deficiencies. Various forms of a test can be developed by taking different samples from the domain of test items conforming to the universe specifications.

The controversy over step 5 points to a popular misconception about the nature of criterion-referenced testing. Some people believe that a test for this purpose is essentially one which classifies an examinee as competent or incompetent with respect to some skill. This is not widely viewed by test theorists as a necessary characteristic of such a test, although some theorists view it as highly desirable for practical applications. What is necessary is that the score be directly interpretable in terms of behaviors or performances. Deciding what level of performance constitutes competence is an extra step.

At least some of the procedures outlined here can be helpful even where a teacher simply wants a mid-term exam that will only be used once. At least stating the rationale and specifying crucial behaviors (either dispositions or performances) can be helpful in thinking through the test specifications.

Test format. An informal logic teacher interested in constructing an achievement test is faced with the problem of deciding what type of item or items to construct, e.g., multiple-choice or essay. There is an amazing variety of item forms used in tests, but for discussion purposes, we will classify them into three main types: multiple-choice, short-answer, and essay. These types are distinguished by the latitude a student has in constructing an answer. Multiple-choice tests allow a student to pick only from

specified possible answers. There are many kinds of multiple-choice forms available (true-false, matching, multiple-response, etc.) and many ingenious variations have been invented to measure a wide range of objectives (see Anderson, 1972, and Wesman, 1971). The short-answer item, such as a sentence-completion item or identification question, allows students more freedom in that they must supply the answer themselves. Students are limited to some extent by the space allowed for the answer and the necessarily limited nature of the question asked. The essay or open-ended answer allows students a great deal of freedom in choosing what they believe to be a good answer.

The problem of the type of item to employ is a thorny one. Test theorists have traditionally favored multiple-choice items, the type for which the concept, universe of behaviors, is best adapted. Such items are easily and inexpensively scored and are not susceptible to errors of measurement caused by inter-grader disagreement. (But comparable errors slip in through the writings of the item and the directions--which always leave room for differing interpretations by examinees.) The data produced by such tests can be analyzed by the sophisticated statistical techniques available to test theorists. On the other hand, students might recognize a multiple-choice alternative as correct when they would not have been able to recall the answer had they been asked a short-answer question.

Essay questions require even more effort and ability on the examinee's part since the structure and content of the answer must be supplied by the examinee. In many cases, informal logicians will want to assess the type of knowledge that essay questions seem best suited to assess. Unfortunately, the concept, universe of behaviors, is especially problematic with essay questions.

Since the type of item one chooses to construct depends on the type of knowledge one is trying to assess, it could be useful to have some classification scheme for types of objectives and "behaviors". Benjamin Bloom and his associates have developed a popular scheme for classifying cognitive educational objectives (Bloom, Engelhart, Furst, Hill, & Krathwohl, 1956). The hierarchical list of terms developed by Bloom, et al., is as follows: knowledge, comprehension, application, analysis, synthesis, and evaluation. Although widely used in the literature on education, this list embodies a host of philosophical and other problems. They are in the list itself and its application to testing. The simple problem of classifying an objective (say, "ability to identify and unstated assumption") gives one doubts about this list. Bloom, et al., actually classify this objective under 4.10, "Analysis of Elements", but it is not clear why they place it there instead of somewhere else.

Although the traditional test-theory view has been that multiple-choice items are to be preferred whenever possible, some writers have recently proposed that, for epistemological reasons, multiple-choice items are the least desirable type. Hugh Petrie (in press) has argued that we should view a test as the introduction of a disturbance that the examinee will correct if the desired achievement has been attained. By limiting the possible responses to a test item (the disturbance), we limit possible novel responses that would also counteract the disturbance. New theories proposed by Petrie and others will no doubt have an influence on the future of testing and evaluation.

Using Tests and Other Techniques in Evaluating Informal Logic Students, Courses, and Curricula

Even the most well-constructed test is not worth much if it is not used properly. A test may give us some information about the performance level of our students or the effect on test scores of different approaches to teaching informal logic. But we do not give tests just to obtain such information. We use the information to make judgments about student achievement or the merit of innovative teaching methods. When we do this we are engaged in the process of evaluation.

There are many different things that we evaluate in education. They cannot all be evaluated the same way. Even in any one particular area, there is not universal agreement about how evaluations should be carried out. Nevertheless, we can offer some advice on the use of tests and other techniques for the purpose of evaluation.

Testing and Evaluation

Tests are certainly the most widely used instruments in evaluation studies. Although there are many legitimate uses of tests, they can also be misused. A carefully prepared plan for an evaluation study can help guard against the misuse of tests. We shall discuss the use of tests to assess student performance and the employment of tests in various experimental designs. Tests can also be used for other purposes (e.g., placement), but the following topics will probably be of greater interest to those interested in teaching and research in informal logic.

The use of criterion-referenced testing in evaluation. Criterion-referenced testing is the assessing of a student's mastery. We cannot endorse the current attempt to put this purpose in terms of a universe of behaviors, but feel that no matter how we conceptualize the basis, criterion-referenced testing of some sort is useful.

One may use the test results to assign grades; or to determine which students should advance to the next unit of study and which should remain behind for more work. In either case, one must face the problem of specifying what level of performance indicates mastery of the material studied (or what levels of performance correspond to a certain degree of mastery). Experts in this area of test theory have no help to offer us on this problem, but it is essential that the person who does decide be thoroughly familiar with the test rationale, its items, and the subject matter of informal logic.

The use of norm-referenced testing in evaluation. If one is interested in comparing the relative standing of groups (e.g., informal logic class vs. traditional class) on some variable (e.g., critical thinking ability), then one employs norm-referenced testing. In selecting a test one must determine whether the face or content validity and construct validity of a test under consideration for use as a measuring instrument are appropriate for one's own purpose. It is here that a list of written course objectives would be helpful, for the test specifications for a particular informal logic test might not match the course objectives. A test may, however, measure some things considered important in the course specification and could therefore serve as a partial measure of the constructs under investigation. Unfortunately, people all too often rely exclusively on the title of a test for information about what the test measures. There is no substitute for a careful, critical examination of a test and its manual.

Experimental design. Just as important as choosing a measuring instrument is the choice of an experimental design. Even the best instruments cannot produce useable data unless a proper testing schedule is followed. We will briefly consider some of the more popular designs. The

reader is encouraged to examine more thorough treatments of this topic, such as, Campbell & Stanley (1963), Winer (1962), or Wiersma (1975).

There is one "design" which most experts do not consider a design at all. In this "design", the results of a pretest and a posttest on an experimental group are compared. (A pretest is a test administered before a treatment. A posttest is a test administered after a treatment. A treatment is any deliberately-introduced change in the environment of the group under investigation. For example, instruction in informal logic would be a treatment.) The problem here is that even if the posttest scores are higher than the pretest scores, one has no reason to attribute this inference to the treatment. Any number of other factors (e.g., maturation, familiarity with the subject-matter induced by the pretest, etc.) could be responsible for the higher scores.

In order to draw meaningful conclusions from the preceding experiment, we also need a control group with which to compare the experimental group. By "control group" we mean a group that is supposed to have the same characteristics as the experimental group except that it does not receive the treatment. The preferred method for obtaining equivalent groups is to choose both groups randomly from the population under study. (There is some disagreement among theorists about whether randomly selected groups are equivalent by definition or whether they are only highly likely to be equivalent. The former position seems to be the commonly-accepted view. Thus this concept of equivalent groups employed by test theorists and statisticians differs from the ordinary concept.)

The simplest type of experimental design is the posttest-only control group design in which a posttest is administered to the control group and

experimental group. One then looks for a significant difference between the mean scores of the two groups. This design is simple to set up and is not as widely used as it could be. On the negative side, the statistical tests involved are not as powerful as those used in some other designs, and there is often a lingering suspicion that the groups were not equivalent at the beginning, despite the random-selection process.

The most popular true experimental design is probably the pretest-posttest control-group design: In this set- , both the control and experimental groups are administered a pretest and a posttest. One often-used strategy for analyzing the results is to compute gain scores (the difference between pre- and posttest scores) and to test the difference between mean gain scores for each group for statistical significance. This approach is challenged by some experts, however, (see Cronbach & Furby, 1970). Analysis of covariance is now one of the recommended procedures for analyses of data from this design. Roughly speaking, analysis of covariance attempts to compare posttest scores while statistically holding pretest scores (and perhaps other variables) constant. This design gives one a check on group equivalence through a comparison of pretest scores. However, it requires twice as many test administrations as the posttest-only control-group design and there is a problem with attempting to generalize to the unpretested population.

One big stumbling block to the use of true experimental designs is the difficulty in arranging for random assignment of individuals to groups. Most institutional settings are not flexible enough to permit randomization, and, in some cases, there are also ethical and political problems with this manipulation of subjects. In such cases, which in educational institutions is most cases, researchers turn to quasi-experimental designs,

which are similar to true experimental designs, but differ in a few important respects.

The most widely used quasi-experimental design is the "nonequivalent" control-group design. This design resembles the pretest-posttest control-group design except that subjects are not assigned to groups at random. The groups are taken as they are found in some institutional setting. This means that extraneous factors that influence the selection process may turn out to be responsible for any significant differences which are found. This possibility must be carefully considered when weighing the evidence collected. If one has information about relevant characteristics of the subjects, this can sometimes be taken into account in the statistical treatment of the data by means of techniques such as analysis of covariance. Since this design is often the only one available, readers interested in research that will be conducted under conditions that preclude the use of true experimental designs should consult more detailed treatments of this and other quasi-experimental designs (e.g., Campbell & Stanley, 1963, Kerlinger, 1964, or Airasian, 1974).

Regardless of the type of experimental design chosen, one problem that any researcher must face is generalizing from a sample to a population. Unfortunately, the populations from which samples are drawn in educational research efforts are almost never the populations over which it would be desirable to generalize. For example, one might draw random samples (for a control group and an experimental group) from all the students taking informal logic during a particular semester for the purpose of evaluation of a certain method of teaching logic. The population to which one would like to generalize is that of all informal logic students, including next year's group. But the population from which the sample was drawn was much

more restricted. (Each member of a population must have an equal chance of being selected in a random assignment.) Sometimes arguments are offered for the typicality of groups chosen in an attempt to generalize over a larger population (Campbell & Stanley, 1963, discuss this difficulty, calling it the problem of "external validity").

Most commercially-available tests contain tables of norms with which one can compare experimental groups or individuals. Such comparisons are useful for suggesting hypotheses about differences between, e.g., national norms and locally-collected data. They can also give an individual an idea of how he or she stands compared to a norm group. The more accurately the norm groups are described, the more readily one can choose the appropriate comparison group. One should not, however, view normative data as a substitute for control-group data collected by the experimenter.

Statistical significance. In educational research, a result that, given the assumptions, is the sort that could have occurred by chance less than five (sometimes one) times out of a hundred is generally deemed to be statistically significant (this is known as the .05 level of significance). Beware of this approach. With very large groups statistical significance can be attributed to differences that are for practical purposes very small. It is therefore good practice to ask about statistically-significant differences whether they are also practically-significant as well. This requires that one immerse oneself in the situation and inquire about the economic and human cost of producing a given difference, and about whether the difference produced is large enough to be concerned about.

Other Approaches to Evaluation

Thus far we have discussed tests and their use as evaluation instruments. The devoting of a large proportion of this paper to tests and test theory

reflects the extent to which this approach to evaluation dominates education at the present time. Are there any other approaches to evaluation?

The answer to the preceding question depends on the extent to which the testing model can be extended to cover everything to be evaluated by educators in general and informal logicians in particular. According to some test experts, most of the efforts expended in evaluation projects should be directed toward the construction and perfection of good tests. For them, evaluation means measurement, and measurement means the use of some instrument covered by some test-theory model.

Responsive evaluation. For some evaluators, however, testing is not the whole of evaluation or even the most important part. One group that employs a somewhat different approach to the evaluation of educational programs and materials is the Center for Instructional Resources and Curriculum Evaluation (CIRCE) at the University of Illinois at Urbana-Champaign, directed by Robert E. Stake. Stake is suspicious of traditional evaluation methods, since they are what he calls "pre-ordinate" (Stake, 1967, 1976, Stake & Hoke, 1976). That is, they depend on prespecified notions of how a successful program or course must appear and be. By looking only for certain kinds of results (usually in terms of test scores), traditional evaluators may overlook things that would be considered just as valuable as the prespecified objectives, if they were noticed.

CIRCE evaluations tend to include a great deal of narrative or "portrayal" material gathered by observers. These observers make note of things they feel are important and judgments of students, teachers, parents, and school administrators. Rather than evaluating a program or course strictly in

terms of test scores, Stake's "responsive evaluation" tries to employ a more holistic approach.

Semi-structured evaluation. Although many of Stake's criticisms of traditional evaluation methods are certainly to be needed, it is by no means clear that tests should be abandoned or even demoted in importance. Rather we feel that evaluators must become more cautious in their interpretations of test results and must become more flexible in their use of other approaches to evaluation (e.g., by including the reports of trained classroom observer in evaluations). The need for flexibility and new evaluation methods is especially pressing in research in informal logic. The Illinois Rational Thinking Project is examining several methods for evaluating curriculum materials in critical thinking. We have found that tests alone do not provide all the information we would like, although we still consider them an indispensible part of evaluation. We are beginning to examine other evaluation methods, some of which are indicated below. Whether these techniques will prove useful remains to be seen, but we invite others to experiment with them and hope others interested in informal logic will make additional suggestions.

Many skills that informal logicians wish to teach their students are not amenable to evaluation by means of traditional tests. For example, the application of informal logic skills in conversation and in everyday arguments is an extremely complicated process. By observing human interactions that are more or less structured, one can begin to get a feel for students' abilities in this area. On the more structured side, debates provide a format that might even produce quantitative data if some type of scoring system is employed. Students must both construct and criticize arguments in a debate, so this particular activity is one which teachers of informal logic should consider using in their classes. Scoring procedures

need to be developed by informal logicians, since what they perceive as good and bad in a debate is different from what the rhetorician sees as good and bad.

Debates, while useful in the evaluation of instruction are not very realistic forums for the application of logical skills. One problem with them is that they do not allow participants to change their positions when they hear a good argument from an opponent (see Scriven, 1976). Small group discussions might provide a more realistic setting for the application of logical skills in a context likely to be found in everyday life. An interview situation might also provide a good context in which to evaluate the ability of students to construct and criticize arguments. Like debates, discussions and interviews might lend themselves to analysis by means of a scoring system, especially if the topic is one in which certain lines of argument could be expected. However, remembering Stake's criticisms of traditional methods, one should not rely exclusively on a scoring key when evaluating something as open-ended as a discussion or interview. An evaluator must be able to spot unforeseen moves that would indicate that students are employing the skills and concepts that have been taught.

Surveys and questionnaires. While surveys and questionnaires are used to some extent at the present time, they are not being employed as fully as they could be. At the primary and secondary level how a course is perceived by other teachers, parents, administrators, and, especially, the students themselves can be important factors in the success or failure of a course. At the college level, how the course is perceived by students is still a very important factor. Whether students view a course as training in the rational pursuit of truth or as training in sophistry will certainly affect our evaluation of the course. Some attempt is now made to analyze

surveys on the basis of traditional test theory models, but it is not at all clear that that model is appropriate. More investigation is needed in this area.

Long-term follow-up. Probably the most neglected approach to evaluation is the long-term follow-up study. While this approach might fit under traditional testing models, this depends on the type of follow-up performed. Unfortunately these kinds of studies are rarely done. This is a rather sorry state of affairs since the effects of most educational programs are meant to be lasting. However, most programs and courses are evaluated at the end of the treatment period and follow-up studies are very expensive and difficult. Lindquist (1951) distinguishes betweem immediate objectives, those which end-of-course evaluations measure, and ultimate objectives, the attainment of which can perhaps only be evaluated at some time long after the treatment period. We suspect that informal logicians will be especially concerned with ultimate objectives, since informal logic courses are meant to help people reason in everyday situations throughout life. Without long-term follow-up-studies, it is difficult to see how one could decide whether ultimate objectives had been attained. The financial and logistical problems which are inherent in long-term studies are obvious, but this fact does not reduce the need for such studies. Rather it counsels us to be well-prepared (and supported) before venturing such a study.

Summary

We have presented a brief overview of the state of testing and evaluation as applied to informal logic.

Currently available general tests in this area were described and criticized. They are the Watson-Glaser Critical Thinking Appraisal, the

Cornell Critical Thinking Test, Level X, the Cornell Critical Thinking Test, Level Z, and (if one groups them together) the Instructional Objectives Exchange Indexes.

Two standard, generally-desirable characteristics of tests were explained: reliability, the tendency of a test to give the same result when given again in the same circumstances; and validity, the characteristic of measuring (or appraising) what the test is supposed to measure (or appraise). Test-retest, parallel form, and internal consistency methods of estimating reliability were described and criticized, and the danger of using multiple internal-consistency methods for tests of heterogeneous traits was noted. Five common approaches to validity were considered: face, content, construct, predictive, and concurrent. Current test-specialist contempt for face validity was questioned; the notion of content validity was challenged because of its intimate relation to the problematic concept, universe of behaviors; construct validity, the idea that a test is valid to the extent that its results fit into a good theory, was explored and found vague, but not uselessly so; and predictive and concurrent validity were deemed to be generally of little use to informal logicians because of the lack of an outside criterion to validate informal logic tests.

We distinguished criterion-referenced testing from norm-referenced testing. In doing so, we suggested a shift in testing-theory vocabulary from "test" to "testing", the reason being that a test developed for one purpose could conceivably be used for the other. Criterion-referenced testing has the purpose of assessing degree of mastery; norm-referenced testing has the purpose of discriminating between and among students and groups. Norm-referenced testing theory is well-developed, though there are

problems, including the built-in invitation to develop reliable, invalid tests. Criterion-referenced testing theory is in its infancy, and in particular has problems with its lack of guidelines for determining a level that shall be deemed mastery and with its generally-accompanying concept, universe of behaviors. It is not clear whethe the recommended random sample from the universe is to be taken of behaviors as dispositions, of behaviors as performances, or of items.

Some procedures for developing one's own informal logic tests were suggested, and various types of evaluation instruments (in addition to the heavily-emphasized multiple-choice tests) were described and recommended.

Experimental designs were considered. We do not recommend the simple pretest-posttest design unless there is a control group. But even if one has a control group, experimental theory calls for the random selection of the subjects for the experimental and control groups from the population about which we want to draw conclusions. This is impossible if we want to draw conclusions about next year's classes, for example, so compromises are struck. One compromise is to draw one's initial conclusions only about the group from which one did manage to draw a random sample, and then attempt somehow to infer to the larger group on the basis of its typicality. A second compromise that is often struck is to pick one's experimental and control groups not at random, but so that they are as comparable as we can get them, and then to assume that they are comparable enough, or to use statistical techniques that, it is hoped, compensate for incomparability (this is called a "quasi-experimental design"). There is no perfect resolution of these problems.

As we proceeded in laying out this introductory treatment of testing and evaluating in informal logic, we broached a number of philosophical problems

that are embedded in this field. We did not attempt an exhaustive list of such problems, but did allude to the following: What sense can be made of random sampling from a universe of behaviors? What is a "behavior"? What is a true score? What is critical thinking? What is rational thinking? What is informal logic? What is the relationship between test performance and mental traits? What is mastery and in general how can mastery be inferred from test performance? Is it plausible to judge a test to be valid on the ground that it fits into a well-confirmed theory, as is recommended by the construct-validity approach? If so, then what rules and procedures can be followed to make such judgments? What constitutes typicality? Can one specify guidelines for generalizing beyond a population from which a random sample was drawn? If so, what are they? Can one specify guidelines for acceptable alternatives to random sampling? If so, what are they?

We mention these problems partly in order to warn interested informal logicians that the field of testing and evaluation is not out there all ready to provide a neat, clean service to us. But we do so also in the hope that some philosophers will undertake work on these or other evaluation-related problems with the intention of offering theoretical help in this area. In view of informal logicians' practical interests in evaluating informal logic competence, it should be apparent that philosophical work on these problems would be a socially-significant activity. We also feel that such work is intrinsically interesting and philosophically important.

We also hope that other informal logicians will develop various kinds of instruments for evaluating informal logic competence. More are needed, and if we do not do it, someone else will--someone who knows even less about it than we do.

References

Ahman, J. S., & Glock, M. D. Evaluating pupil growth. Boston: Allyn and Bacon, 1958.

Airasian, P. W. Designing summative evaluation studies at the local level. In W. J. Popham (ED.), Evaluation in Education. Berkeley, Calif.: McCutchan, 1974. Discussion of choosing an experimental design under the constraints imposed by typical institutional settings.

American Psychological Association. Standards for educational and psychological tests. Washington, D.C.: American Psychological Association, 1974.

Anderson, R. C. How to construct achievement tests to assess comprehension. Review of Educational Research, 1972, 42, 145-170. Contains practical suggestions for writing types of items which are useful in criterion-referenced tests.

Bechtoldt, H. P. Construct validity: A critique. American Psychologist, 1959, 14.

Bloom, B. S., Englehart, M. D., Furst, E. J., Hill. W. H., & Krathwohl, D. R. Taxonomy of educational objectives, handbook I: Cognitive domain. New York: David McKay, 1956.

Buros, O. K. (Ed.) The seventh mental measurements yearbook (2 vols.). Highland Park, N.J.: Gryphen Press, 1972. Standard sourcebook for psychological tests.

Campbell, D. T. & Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 1959, 56, 81-105.

Campbell, D. T. & Stanley, J. C. Experimental and quasi-experimental designs for research on teaching. In N. C. Gage (Ed.), Handbook of research on teaching. Chicago: Rand McNally, 1963. Classic article on experimental designs for educational research.

Cronbach, L. J. Validation of education measures. In P. H. DuBois (Ed.), Proceedings of the 1969 invitational conference on testing problems. Princeton, N.J.: Educational Testing Service, 1969. Preliminary version of Cronbach, 1971.

Cronbach, L. J. Test validation. In R. L. Thorndike (Ed.), Educational measurement. Washington, D.C.: American Council on Education, 1971. A seminal paper on test validity. Some parts are technical, but it is recommended reading, nonetheless.

Cronbach, L. J. & Furby, L. How we should measure "change"--or should we? Psychological Bulletin, 1970, 74, 68-80. Arguments against the use of gain scores and the pretest-posttest control group experimental design.

Cureton, E. E. Reliability and validity: basic assumptions and experimental designs. Educational and Psychological Measurement, 1965, 25, 327-346.

Ennis, R. H. An appraisal of the Watson-Glaser critical thinking appraisal. Journal of Educational Research, 1958, 52, 155-158. Covers earlier versions of the Watson-Glaser test (Forms Am and Bm).

Ennis, R. H. Assumption-finding. In B. O. Smith & R. H. Ennis (Eds.), Language and concepts in education. Chicago: Rand McNally, 1961.

Ennis, R. H. A concept of critical thinking. Harvard Educational Review, 1962, 32, 81-111. With a few minor amendments, this notion of critical thinking is the basis for the Cornell Critical Thinking Tests.

Ennis, R. H. Operational definitions. American Educational Research Journal, 1964, 1, 183-201.

Ennis, R. H., Gardiner, W. L., Morrow, R., Paulus, D., & Ringel, L. The Cornell Class Reasoning Test. Urbana, Ill.: Illinois Critical Thinking Project, 1964.

Ennis, R. H., Gardiner, W. L., Morrow, R., Paulus, D., Ringel, L., & Guzzetta, J. The Cornell Conditional Reasoning Test. Urbana, Ill.: Illinois Critical Thinking Project, 1964.

Ennis, R. H., & Millman, J. Manual for Cornell Critical Thinking Test, Level X and Cornell Critical Thinking Test, Level Z. Urbana, Ill.: Illinois Critical Thinking Project, 1971(a).

Ennis, R. H., & Millman, J. The Cornell Critical Thinking Test, Level X. Urbana, Ill.: Illinois Critical Thinking Project, 1971(b).

Ennis, R. H., & Millman, J. The Cornell Critical Thinking Test, Level Z. Urbana, Ill.: Illinois Critical Thinking Project, 1971(c).

Ennis, R. H., & Paulus, D. Critical thinking readiness in grades 1-12 (phase I: deductive logic in adolescence). Ithaca, N.Y.: Cornell University, 1965. (ERIC Document Reproduction Service No. ED 003 818).

Glaser, R., Klaus, D. J. Proficiency measurement: Assessing human performance. In R. M. Gagne (Ed.), Psychological principles in systems development. New York: Holt, Rinehart, and Winston, 1962.

Glass, G. V. Standards and criteria. Journal of Educational Measurement, 1978, 15, 237-261.

Guilford, J. P., & Hertzka, A. F. Logical Reasoning (test). Orange, Calif.: Sheridan Psychological Services, 1955.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. Criterion-referenced testing and measurement: a review of technical issues and developments. Review of Educational Research, 1978, 48, 1-47. A review of the state-of-the-art in criterion-referenced testing. Some technical sections.

Haney, W. V. The uncritical inference test. Wilmette, Ill.: William V. Haney Associates, 1975.

Hempel, C. G. A logical appraisal of operationism. In P. G. Frank (Ed.), The validation of scientific theories. New York: Collier, 1961.

Hempel, C. G. Aspects of scientific explanation. New York: Free Press, 1965.

Hempel, C. G. Philosophy of the natural sciences. Englewood Cliffs, N.J.: Prentice-Hall, 1966.

Instructional Objectives Exchange. Judgment: deductive logic and assumption recognition, grades 7-12. Los Angeles: Instructional Objectives Exchange, 1971.

Kelley, T. L. Interpretation of Educational Measures. Yonkers, N.Y.: World Book Co., 1927.

Kerlinger, F. N. Foundations of behavioral research. New York: Holt, Rinehart, and Winston, 1964. In-depth discussions of experimental designs.

Lindquist, E. F. Some preliminary considerations in objective test construction. In E. F. Lindquist (Ed.), Educational measurement. Washington, D.C.: American Council on Education, 1951.

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968. The definitive work on norm-referenced test theory.

Messick, S. The standard problem: meaning and values in measurement and evaluation. American Psychologist, 1975, 30, 955-966.

Petrie, H. Against objective tests: a note on the epistemology underlying current testing dogma. In Mark Ozer (Ed.), Toward the more human use of human beings: Issues in the application of cybernetics to assessment of children (forthcoming).

Popham. W. J. (Ed.). Criterion-referenced measurement. Englewood Cliffs, N.J.: Educational Technology Publications, 1971. Good introduction to criterion-referenced testing.

Sax, G. The use of standardized tests in evaluation. In W. J. Popham (Ed.), Evaluation in education. Berkeley, Calif.: McCutchan, 1974. Contains a comparison of criterion-referenced and norm-referenced tests.

Scriven, M. Reasoning. New York: McGraw-Hill, 1976.

Sell. D. E. Evaluation aptitude test-manual. Munster, Ind.: Psychometric Affiliates, 1952.

Smith, R. A. Regaining educational leadership: Critical essays on PBTE/CBTE, behavioral objectives, and accountability. New York: John Wiley, 1975.

Stake, R. E. The countenance of educational evaluation. Teacher's College Record, 1967, 68, 523-540.

Stake, R. E. To evaluate an arts program. Journal of Aesthetic Education, 1976 10, 115-133.

Stake, R. E., & Hoke, G. A. Movement and dance in a downstate district. The National Elementary Principal, 1976, 55.

Stanley, J. C. Reliability. In R. L. Thorndike (Ed.), Educational measurement (2nd ed.). Washington, D.C.: American Council on Education, 1971. A thorough treatment of the topic. Somewhat technical.

Stewart, B. Testing for critical thinking: A review of the resources. Urbana, Ill.: Illinois Critical Thinking Project, 1979.

Strawson, P. F. Introduction to logical theory. London: Methuen & Co., 1952.

Watson, G., & Glaser, E. M. Manual for Watson-Glaser critical thinking appraisal. New York: Harcourt, Brace, & World, 1964(a).

Watson, G., & Glaser, E. M. Watson-Glaser critical thinking appraisal, form Ym. New York: Harcourt, Brace, & World, 1964(b).

Watson, G., & Glaser, E. M. Watson-Glaser critical thinking appraisal, form Zm. New York: Harcourt, Brace, & World, 1964(c).

Wesman, A. G. Writing the test item. In R. L. Thorndike (Ed.), Educational measurement. Washington, D.C.: American Council on Education, 1971.

Wiersma, W. Research methods in education (2nd ed.). Itasca, Ill.: F. E. Peacock, 1975.

Winer, B. J. Statistical principles in experimental design (2nd ed.). New York: McGraw-Hill, 1971.

Additional Readings

Freedman, D., Pisani, R., & Purves, R. Statistics. New York: Norton, 1978. Introductory text.

Gage, N. L. (Ed.). Handbook of research on teaching. Chicago: Rand McNally, 1963.

Glaser, R. & Nitko, A. J. Measurement in learning and instruction. In R. L. Thorndike (Ed.), Educational measurement. Washington, D.C.: American Council on Education, 1971. Contains a discussion of some uses of criterion-referenced tests.

Glass, G., & Stanley, J. C. Statistical methods in education and psychology. Englewood Cliffs, N.J.: Prentice-Hall, 1970. Widely used text, moderate level of difficulty.

Millman, J. Criterion-referenced measurement: Current applications. In W. J. Popham (Ed.), Evaluation in education. Berkeley, Calif.: McCutchan, 1974. A thorough introduction to criterion-referenced testing.

Scriven, M. The methodology of evaluation. In R. W. Taylor, R. M. Gagne, & M. Scriven, Perspectives in curriculum evaluation. Chicago: Rand McNally, 1967. Scriven here introduces an important distinction between formative and summative evaluation.