# articles

# Problems in Testing Informal Logic Critical Thinking Reasoning Ability[1]

**Robert H. Ennis**
**University of Illinois**

To my knowledge there are now five English-language machine-gradeable tests readily available on the North American Continent that are, or might be construed as, critical thinking tests: the two Cornell critical thinking tests (Ennis & Millman, 1982a, 1982b); **New Jersey Test of Reasoning Skills** (Shipman, 1983); **Ross Test of Higher Cognitive Processes** (Ross & Ross, 1976); and **Watson-Glaser Critical Thinking Appraisal** (Watson & Glaser, 1980).[2] But there will be more critical thinking tests, because of the greatly increased emphasis critical thinking and informal logic are receiving these days. More instructors will want a quick way to compare their students with others, and will turn to machine-gradeable tests. Furthermore many of us are engaged in discussion (sometimes dispute) over what gets taught under a label like "critical thinking", and how long it continues to be taught. One instrument in such discussion or dispute is the machine-gradeable multiple-choice critical thinking test.

Since I use the terms "critical thinking" and "informal logic" and "reasoning" roughly interchangeably as labels for an area of concern, I economize here by using only the term, "critical thinking". It is in the names of three of the five cited tests, the terms "reasoning" and "cognitive process" appearing once each. I know of no widely available test containing the term "informal logic" in its title.

It is largely in preparation for this expected increase in critical thinking tests that I attempt in this paper to share my experience in critical thinking testing by noting some problems and by offering possible resolutions. My hopes are 1) that both consumers and developers of critical thinking tests will profit from the sharing; 2) that some members of my audience will help me in my attempts to deal with the problems; and 3) that some even will become deeply enough involved in the problems to work on them. Although the problems are practical, all have philosophical foundations.

The problems with which I shall deal are concerned with the testing for students' value judgments, their induction ability, and their assumption-identification ability; and with the "reliability" (read "'consistency") and validity of critical thinking tests. The problems are broad and varied. I shall not do them full justice.

## VALUE JUDGMENTS

Although making value judgments strikes me as an aspect of critical thinking, I do not think that it is fair for the keying of an answer to depend on a value judgment about which there is possible disagreement, unless the value judgment is constitutive of critical thinking, such as the judgment that it is generally good to be ready and willing to consider openmindedly points of view with which one disagrees.

I realize that the tone of these remarks might suggest a greater clarity and precision for the concepts, **value judgment, openminded,** etc. than they have. But in the context of my comments they seem precise enough. Consider for example two items from a section of **Watson-Glaser Critical Thinking Appraisal,** Form A (Watson & Glaser, 1980), "Test 5: Evaluation of Arguments". The question is whether a strong labor party would promote the general welfare of the people of the United States. For each of the following items the student is to take what is offered as a reason as true, and must decide whether the argument is strong or weak. (To be strong, the reason must be both important and directly related to the question.)

65. No; a strong labor party would make it unattractive for private investors to risk their money in business ventures, thus causing sustained large-scale unemployment.
67. No; labor unions have called strikes in a number of important industries.

Item 65 is keyed "strong"; Item 67 is keyed "weak".

However, a good Marxist might well regard Item 65 as weak on the ground that sustained large-scale unemployment would be a good thing because it would awaken the proletariat. Item 67, on the other hand, might well be regarded as strong by many conservatives who believe that a strong labor party would encourage labor unions, and that strikes in important industries are bad things. It does not seem fair to mark such people wrong in their evaluations of these arguments, so I urge that such items not appear on critical thinking tests.[3] The keying depends on value judgments about which there is possible disagreement and which are not constitutive of critical thinking. The concept **value judgment** seems clear enough in this context for me to make this recommendation.

## INDUCTION

Under the label "induction" I include generalizing from a number of particular instances to a broad statement using the same concepts (for example, inferring from "One railroad tie burned with a foul smell", "Another railroad tie burned with a foul smell", etc. to "Railroad ties burn with a foul smell."). I also include best-explanation inference. Basically the problem is that induction requires background assumptions about the way the world is and works, not all of which can be explicitly specified in a set of directions. A related problem is that people with different levels of sophistication justifiably give different levels of endorsement to a conclusion. Labeling a conclusion "probably true" instead of "true" constitutes a lesser level of endorsement. Saying that the data is insufficient is no endorsement.

Consider Item 6 in the Watson-Glaser test, preceded by a description of a situation.

**Description:**

Mr. Brown, who lives in the town of Salem, was brought before the Salem municipal court for the sixth time in the past month on a charge of keeping his pool hall open after 1 a.m. He again admitted his guilt and was fined the maximum, $500, as in each earlier instance.

**Item:**

6. On some nights it was to Mr. Brown's advantage to keep his pool hall open after 1 a.m., even at the risk of paying a $500 fine.

Since the proposed conclusion is a possible explanation of the facts, this seems to be a case of induction. The choices are "True", "Probably true", "Insufficient data", "Probably false", and "False". The keyed answer is "Probably true."

A very sophisticated person might well adopt the position that we do not know enough about the situation even to say "Probably true." Perhaps he had put his son in charge and thought this was a small price to pay for all the years he had neglected his son. Perhaps in spite of his admission of guilt, he had not kept the pool hall open after 1 a.m., but this was a way to pay off the municipal authorities for granting him a license. If so many possibilities occur to someone, that person, if sophisticated and cautious, might well justifiably decide to mark "Insufficient data".

On the other hand imagine a less sophisticated student who has learned in civics class that people often find it profitable to violate the law and pay the resulting fines, but that fines of that magnitude would deter someone unless it were to the person's advantage to be an offender. Such a student could justifiably mark the item "True". To mark the answer incorrect would be to penalize the student for having empirical beliefs about the way the world works that are different from those of the test authors.

In designing the Cornell critical thinking tests, I attempted to deal with the problem of avoiding the distinction between degrees of endorsement (for example, between "True" and "Probably true"), asking only in which direction, if any, the evidence points, and by seeking topics and items about which I thought it most likely that there would be no significant differences in background knowledge. However, our interviews with respondents have made clear that this program was not totally successful.

For example some of the items in the Cornell Level-X critical thinking test ask about the bearing of certain information on the hypothesis that some missing explorers on the newly-discovered plant, Nicoma, are dead. One piece of information is that the blankets and sheets of the explorers' huts are all found neatly folded in the closets. The intended answer is that this information goes against the hypothesis, because the folding and putting away are not things that would have been done in emergency or disaster. On the other hand, someone with a belief that it is standard practice, even in emergencies, to clean up immediately after the dead and to fold their sheets and blankets, might think that this information neither supports nor goes against the hypothesis. Background beliefs influence the answer here.

Different people do sometimes bring different background assumptions and different levels of sophistication to our induction items. It seems unfair to mark them down for so doing. Accordingly the stance that I have adopted is that we cannot expect 100% agreement with the key on all of these items, but that the best critical thinkers will agree at least 85% of the time.

I do not see this problem as merely a testing problem. It is a problem for anyone who tries to develop a system of rules for judging inductive conclusions. There is always the possibility that something else will turn up that has not yet been figured into the decision. If I am wrong about this, I hope to be so instructed, perhaps so that my "85% stance" can be replaced by a "100% stance".

In his **Critical Thinking and Education,** John McPeck (1981, p. 149) makes some suggestion that seem aimed in part at this induction-testing problem:

> "1. That the test be subject-specific in an area (or areas) of the test taker's experience or preparation. This is required because knowledge and information are necessary ingredients of critical thinking."
> "2. That the answer format permit more than one justifiable answer. Thus an essay might better fit the task, awkward and time consuming as this might be..."
> "3. That good answers are not predicated on being right, in the sense of true, but on the quality of justification given for a response."

McPeck's second and third suggestions call for essay tests that are graded by human experts. I do not see how computers can do it. At the Illinois Thinking Project, Eric Weir and I developed an essay test (Ennis & Weir, 1983) that does call for appraisal of the justification offered. This we feel requires trained appraisers, but it is time consuming, as McPeck suggests. Given an average of six minutes per grading, 5000 tests would take 50 hours. I like the idea of essay tests, but have not found it heavily used.

Furthermore the problem still exists to some extent. Even if the subject matter of the item be within the test-taker's experience there will be differences in the unstated background beliefs of the test-taker and evaluator. The evaluator can make allowances for explicit differences in background beliefs, but not always for implicit ones of which the grader is unaware.

If McPeck means by his first suggestions that critical thinking tests must be in a given subject as taught in schools and colleges, then I must demur. Consider the criterion that a hypothesis is justified only to the extent that plausible alternatives have been ruled out. Not only does this criterion apply very widely (for example, educational research, Shakespearean interpretation), but it applies in areas that are not subjects taught in schools, such as figuring out why there is water in the basement, deciding whether the defendant knew that her act created a strong probability of great bodily harm, and judgining whether Ernie stole the cookies. These last three are enterprises that call for critical thinking and are not subject specific, if one thinks only of subjects as taught in schools. But we do want to teach people how to operate in such areas and we do want to test for competence to do so.

ASSUMPTION IDENTIFICATION

Testing for assumption identification ability faces several problems: 1) a variety of things are called assumptions; 2) assumptions that are significant are not (logically) necessarily made; and 3) the role of background information often makes it unfair to ask **whether** some particular proposition is an assumption.

**The Variety of Things Called Assumptions.**

Often the word "assumption" is a pejorative term, so that in an open-ended test, if asked to find an assumption, a student, unless warned to do otherwise, will usually pick something that the student believes to be dubious, rather than only something that is a crucial support. Furthermore students often pick dubious **conclusions** as assumptions (Doing so is not a violation of standard usage.). There are also Strawsonian presuppositions, unstated gap-filling premises, and unstated back-ups for other premises. (See Ennis, 1982, for further explanation, if these labels do not communicate.) If the test is open-ended, and we do not want conclusions or merely dubious statements, we should say so.

Another choice is between **used** and **needed** assumptions. If the context is such that we want to know what the assumer was actually thinking, we search for used assumptions (assumptions that were actually used, consciously, or perhaps subconsciously, by the thinker). A claim that something is a used assumption is an empirical claim about a mental event, and thus, by my way of thinking, is to be judged on the inference-to-best explanation model. On the other hand, if the context is such that we want to know what the assumer needs to add to the argument to make it least weak, then we look for a needed assumption. If we are trying to decide whether the conclusion is true, we have this sort of context. Here we employ the principle of maximum charity, because we want to give the conclusion its best chance.

In an open-ended test of assumption-identification ability, we should make clear to our students whether the context is one calling for figuring out what the person was thinking, or for figuring out whether to believe the conclusion. Different contexts often call for different assumptions. Multiple-choice critical thinking tests that I know about offer a context in which the purpose is to decide whether to believe the conclusion of an argument for which the assumption is sought. There the basic question is whether the assumption is needed by the argument. If it is not needed, then it would be unfair to attribute the assumption to the argument. This brings us to the second and third problems that I mentioned.

**Logical Necessity.**

As I have argued elsewhere (Ennis, 1982), assumptions that are significant are not logically necessary to an argument. There is always a way around them. Consider this example from the Watson-Glaser test:

> "I'm travelling to South America. I want to be sure that I do not get typhoid fever, so I shall go to my physician and get vaccinated against typhoid fever before I begin my trip."
>
> Proposed assumption:
>
> 28. Typhoid fever is more common in South America than it is where I live.

The key claims that this proposed assumption is "made". The directions say:

> If you think the assumption is not necessarily taken for granted in the statement, blacken the space under "ASSUMPTION NOT MADE."

It is logically possible that typhoid fever is more common where the speaker lives, but that its consequences are more serious if contracted in South America, perhaps because of the climate—or differences in typhoid-care facilities. So the proposed assumption is "not necessarily taken for granted" if the necessity in question is logical necessity. Those students who give such an interpretation to "necessarily" will get this item wrong, and others like it.

But even if the instructions are not given this interpretation, there would be the problem of background information in this and similar items. If I believe the suggested possibility to be a plausible alternative (background information), then again I would be justified in marking "Not made" (contrary to the key) on the basis of my background information.

In order to identify assumptions, whether used or needed, background information is always relevant. Hence it is dangerous to ask in a multiple-choice test whether a particular assumption is made. Rather it seems safer to give a choice of several alternatives, including one and only one gap filler that makes (or easily helps make) a deductively valid argument from the given premise to the given conclusion. This is then the reasonable choice for the answer, so long as it is not less plausible than the other choices (background knowledge sneaking in again), and if the context is one in which the truth of the conclusion is the concern.

Here is an example from **Cornell Critical Thinking Test, Level X** (Ennis & Millman, 1982a):

> 69. "The shorter of the two people wearing green hats is a female. I know because I saw her long hair when she removed her hat." Which is probably taken for granted?
>
> A. All females have long hair.
> B. Only females have long hair.
> C. A person wearing a green hat is likely to be female.

The keyed answer, **B**, makes the argument deductively valid (or does so with minor adjustments if one wants to be strict about it). The word "probably" has been included in the question asked in deference to the fact that full context is not specified, and that background knowledge does matter.

Somewhat in between is the following item from the New Jersey test:

> 8. Josie said, ""This paper must have been written by a boy, because the handwriting is so bad." Josie must be assuming that
>
> a. some boys have poor handwriting.
> b. only boys have poor handwriting.
> c. all boys have poor handwriting.

Although the key is not distributed, presumably the keyed answer is **b.** At least **b** would transform the argument into a

deductively valid one. But a careful thinker might leave it blank on the ground that there is no right answer. That is, Josie does not have be assuming that only boys have poor handwriting, which attribution is actually uncharitable, since it is so obviously false. Josie's argument works if we add the proposition that only the boys in that class group have poor handwriting and that all the papers being considered are from that class group. The lead-in would probably be better stated as follows: "Josie is probably assuming that, in this group...".

In sum I recommend that the type of assumption should be made clear, that we not ask for logically necessary assumptions, that in multiple-choice tests we ask test-takers to choose among several candidates, rather than decide for each whether it is assumed, and that among the choices there be one and only one that contributes readily to the deductive validity of the argument—and that this one not be more inherently implausible than the other choices.

## CONSISTENCY

The testing establishment defines "reliability" as consistency of measurement, and pays much attention to the "reliability" of tests, partly because one can obtain "objective" numbers that indicate consistency, partly because these numbers are generally higher than other numbers one obtains about tests, and partly because it seems like a good idea for a test to be consistent from one administration to the next, though it is somewhat misleading to the public to label consistency in measurement by the term "reliability". The term "reliable" is defined in my Webster's as "trustworthy", which suggests that a reliable test tells us what we want to know, not merely that it gives the same result each time regardless of whether it gives me any information about critical thinking ability, for example. So we could have a reliable test that is called a critical thinking test, according to this technical sense of "reliable", even though it does not test for critical thinking at all. However if we remember that "reliability" in the technical sense means consistency of measurement, not validity, then this problem will not cause trouble.

But the situation is more serious, because the most frequently-used indicators of "reliability" are the Kuder-Richardson formulas, which tell only the degree of internal consistency of a test; that is, the degree to which the items intercorrelate with each other. This is not consistency from one test administration to the next; it is item homogeneity. If **critical thinking** is a heterogeneous concept, then a good comprehensive critical thinking test would probably not do as well on such so-called "reliability" measures as a critical thinking test for only one aspect of critical thinking, say deduction.

There are other indicators of reliability, I should note, including test-retest correlations and correlations between supposedly parallel forms. But at least in part because they are so much easier to use, the Kuder-Richardson formulas are used most often, usually presenting us therefore with a double invitation to misinterpretation.

Item discrimination information generally has the same problem. One would expect from the name that an item discrimination index would tell the extent to which an item discriminates the way it is supposed to discriminate. But the criterion usually used is total score on the test, so average item discrimination indices are generally indicators of internal consistency.

Thus we must remember that several readily-obtainable **apparent** indicators of quality are indicators of internal consistency. The difficult question then is, "How important is internal consistency?" An important part of this question is the question, "To what extent is critical thinking ability a homogeneous ability?" I am puzzled by this question, but my inclination is to say that critical thinking ability is fairly heterogeneous, consisting of such diverse elements as open-mindedness, ability to see other alternatives, experience and background knowledge, knowledge of criteria to apply in thinking critically, ability to handle complexity in an orderly fashion, and some others. All of this is quite speculative. I invite you to join me in the attempt to deal with this question, the answer to which has instructional and curricular implications—in addition to its relevance to the question of how to treat internal-consistency data about critical thinking tests.

## VALIDITY

The problem of determining the validity of a critical thinking test is a difficult one. Standard approaches to validity include criterion-related validity, content validity (old and new), and construct validity. There is discussion of these in **Standard for Educational and Psychological Tests** (Joint Committee, 1974), but there are problems and the booklet is being extensively revised. (The distinction between old and new content validity I introduce here to give greater coherence to the discussion in the light of tradition.) After considering the standard approaches to validity, I shall look at one pesky validity question, "What does the test **really** test?"

### Criterion-Related Validity.

Criterion-related validity is the extent to which the test correlates with an outside pre-established criterion, already accepted as valid. But there really is no outside pre-established criterion for critical thinking ability. I am regretfully suspicious even of teachers' rating of students—even my own ratings of my own students.

### Content Validity, Old and New.

Content validity of the older type depended upon the following of a careful plan to cover the area to be tested, and agreement among experts that the test (with its accompanying answers) does in fact reasonably cover the content. This approach seems the best to me, though securing agreement on anything of this nature is difficult, expecially among philosophers. Needed is agreement about what constitutes critical thinking, about the appropriateness of some particular coverage, and about the answers to the items. All of this is good practical epistemology, so I hope that more philosophers can be persuaded to think of working in the area of critical thinking testing as more than only a fulfillment of their teaching responsibility.

The five tests that I mentioned earlier differ markedly in their content. The three that are actually called "critical thinking" tests (the two Cornell tests and the Watson-Glaser Test) all include sections on deduction, induction, and assumption identification. In adition the Watson-Glaser test includes a section on strong and weak arguments (the one to which I earlier objected because of its testing for a person's value judgments); Cornell Level X has a section on credibility and observation; and Cornell Level Z has sections on credibility, fallacies (especially equivocation), experimental planning and reasoning, and definition.

The New Jersey test (called a "reasoning" test) emphasizes deduction quite heavily, with over half its items on deduction. Assumption identification receives some attention and a variety of other critical thinking aspects are touched upon. It seems to fit the curriculum it was presumably designed to test, that of the Institute for the Advancement of Philosophy for Children,

an advantage or a disadvantage, depending on the extent to which the things emphsized in the curriculum actually reflect critical thinking in a balanced manner.

The Ross test (called a "cognitive processes" test), although it includes sections on deduction and assumption identification, also includes six other sections, some of which one might have trouble calling "critical thinking", for example, a section on verbal analogies.

A sixth test, though its catalogue calls it a critical thinking test, I have not included in my listing, because it contains only deduction items. It is **Logical-Reasoning** (Hertzka & Guilford, 1955). There are other deduction tests available (including some other Cornell tests), but since they do not claim to be critical thinking tests, I shall not discuss them here.

A clear implication of this brief commentary on content is that people vary in their judgments about the appropriate content for a critical thinking test. At least some experts thus are in disagreement, requiring a test consumer to choose among the different conceptualizations of critical thinking.

But there is more. One must not only look at, but look beyond the **names** of the tests and the sections of the tests. One must also look at the items and their keyed answers. For example, the heading, "Strong and Weak Arguments" does not reveal all that is going on in that section of the Watson-Glaser test to which I earlier objected. In the Ross test, the given heading, "Questioning Strategies", fails to reveal that the test-takers do not choose among or devise questioning strategies. Rather they choose among interpretations of information secured by questioning strategies devised by the test authors.

Since critical thinking testing is very difficult, I am not here urging critical-thinking-test consumers to demand perfection. Rather I am urging them to take the trouble to pay close attention to the actual content of an alleged critical thinking test. Although expert opinion is relevant to a content validity judgment of the old type, since the "experts" disagree, a test consumer must look at the content as well as the statistics.

**New.** Content validity (new type) has the appearance of behavioral scientific objectivity, because it calls for random sampling from some universe, but it seems crippled by deep problems, as Thomas Tomko has argued (1981). Often called "criterion-referenced testing" or "domain-referenced testing", its idea is that there is some total universe that is the content. A random sample drawn from this universe should surely be a scientifically-objective representation of the content. The problem is to identify a set of sampleable units that are in fact the content of the field. Candidates for the types of units include "behaviors", responses, test items, and situations. Items, and more broadly, situations that call for responses are at least plausible identifiable and selectable units, but I cannot imagine an exhaustive comprehensive depiction of the content of critical thinking that proceeds by listing such things. There is an infinite number of such situations or possible items. In order to assure a random sample we must provide that each unit in the universe have an equal chance of being selected. I cannot imagine an exhaustive set of critical thinking situations such that one can give each an equal chance of being selected. Hence new-type content validity seems an inappropriate approach for judging critical thinking tests.

**Construct Validity.**

The theory of the third type of validity, construct validity, is still being developed (see Cronbach, 1971; Norris, 1981). The motivation is the perceived difficulties with the other types of validity, in particular the lack of a pre-established outside criterion with which to correlate the results of a test under investigation. Roughly speaking, the idea here is that a test is justifiably believed valid to the extent that information about it fits with other information we have. This is a vague notion, ripe for sharpening and investigation by philosophers of science, as Norris is doing. However, regardless of the outcome of the investigation, it will at least for a long time be difficult to claim for any critical thinking test that it is valid in this way, because of the looseness of the concept **critical thinking**, and because our scientific knowledge about the human activity of critical thinking, is at the air-earth-fire-water stage, and perhaps will be there for a long time.

In sum, those who develop critical thinking tests will find it difficult to make a convincing case for their validity. Describing the structure of the test and inviting people, including experts, to look it over seems like the best approach now. Correspondingly a person trying to determine whether a test is valid should be cautious in judging scientific-appearing claims for validity, and should look carefully at the items and the proposed answers, paying heed to the structure and basis of construction of the test.

**"What does the test really test?"**

Often a claim is made that a test really tests for something other than what its name suggests. Earlier I was implicitly suggesting that for some people Test 5 of the Watson-Glaser test really tests in part for their values. McPeck (1981, p. 146) suggested that the induction items in **Level Z** of the Cornell critical thinking tests (Ennis & Millman, 1982b) "are clearly questions of reading comprehension more than anything else".

Another thing that critical thinking tests are claimed to really be testing is general intelligence, on the ground that they correlate substantially with intelligence tests. Michael Scriven (personal communication) ascribed such a claim to significant figures in Educational Testing Service. McPeck (1981, p. 142) made such a claim about the Watson-Glaser test.

Are these "really-tests-for" claims testable? If such claims are, as I think, responsibility-ascribing **casual** claims, McPeck's claim about reading might then be translatable into the following: "The cause of significant variation in **Level Z** scores among the members of the population being tested is variation in reading ability." (The reduction, based on correlation, of critical thinking to intelligence in addition seems to assume a strongly positivist principle of parsimony and reduction.) If I am correct about the causal part of my suggestion, then by my non-reductive analysis (Ennis, 1973) of effect-explaining causal statements, the McPeck reading claim is that variations in reading ability 1) are sufficient, given the circumstances, to produce the variations we get in test scores, and 2) are responsible for them. If this is so, one prediction might be that there would be high correlations between reading scores and induction items at all levels of the population in question. Another prediction is that attempts by informal logicians to teach induction skill to college students would fail to produce improvement in their **Level Z** induction scores, unless the instruction is instruction in reading, or at least improves their reading ability. These predictions suggest at least partial testability of McPeck's claim.

I am trying to do several things here: to suggest a way of understanding the charge that a test really tests for something else, to warn critical thinking testers that the charge might well be leveled at them, to suggest ways of responding to the charge, and to show again the relevance of traditional

philosophical concerns (in this case concerns with causation and testability) for the informal logic movement.

**SUMMARY**

In my remarks, I have covered much ground, but have neglected many possible refinements. I have tried to share my experience in facing the practical and philosophical dimensions of some critical thinking testing problems in the hope that this would be of help to consumers and developers of critical thinking tests, in the hope that my audience could help me deal with these problems, and in the hope that more specialists of many sorts, including philosophers, would devote their talents to these practical problems.

I have broken the problems into two groups: critical thinking content (value judgments, induction, and assumption identification) and testing concerns (internal consistency and validity), but remember that one of the testing concerns is critical thinking content. These are not the only problems; I have picked some that have particularly stimulated me.

In the area of value judgments I suggested that we not allow a student's score to depend on the student's agreement with our value judgments in controversial areas—except for values constitutive of critical thinking. In the area of induction I claimed the dependence of judgments to some extent on background beliefs and level of sophistication, and suggested that we try to seek items that required background beliefs on which there would be heavy agreement, and that we not ask students to make the distinction among degrees of endorsement ("True", "Probably true", etc.). I do not recommend that all critical thinking testing should be in specific subjects as taught in the schools and colleges, because so much critical thinking in real life is not thus artifically delimited.

In the area of assumption identification I recommended that when asking for open-ended assumption identification we be clear about the kind of assumptions we are seeking, and that we not ask for logically-necessary assumptions; and that for multiple-choice testing we not ask whether an assumption is made, but rather ask which of several candidates is probably assumed, given the choices and given some situation. I also suggested that one acceptable item-type would have as one and only one of its choices a statement that would fill the gap in (or best help to fill the gap in) a deductive argument.

I have focused on multiple-choice tests, because they have certain practical advantages, but I do think that some of the problems I mentioned can be handled by essay testing—with grading by people who are good at critical thinking and are flexible enough to adjust their scoring to accommodate good arguments and insights that are different from those expected.

In the area of internal consistency of tests, I noted the use of the word "reliability" to refer to consistency of repeated measure and the use of internal consistency as an indicator of this "reliability". A problem here is the extent to which **critical thinking** is a homogeneous concept. I warily suggest that it is not.

In the area of validity, I suggested the inapplicability of criterion-related validity and new-type content validity, and the difficulty of application of old-type content validity and construct validity, but did suggest initial emphasis on old-type content validity, and warned of the differences in content among existing tests. I also suggested a causal interpretation of claims that a test really tests for something else, and interpreted such claims in terms of a non-reductive responsibility analysis of effect-explaining causal claims. From this I suggested some possible preditions that might be generated 1) to test claims about what a test really tests, and 2) to explore the testability and meaning of such claims.

Some philosophical questions that are foundational here include the following:

What is critical thinking?

In what way are value judgments different from empirical judgment?

Can there be rules for induction, the application of which does not depend on unspecifiable outside knowledge?

Is "probably" a degree-of-endorsement specifier?

What is the role of deduction in real arguments?

How do you tell what is assumed?

Is critical thinking ability a homogeneous trait?

How does one judge the fittingness of a test into an array of information and beliefs?

What do effect-explaining causal claims mean?

What constitutes a check on testability?

I have not tried to discourage you by suggesting these difficulties and questions. The situation, although imperfect, is not a disaster. Actually, I believe that any one of the five tests listed is worth using and could be quite helpful. In a perverse way, I am trying to encourage by provocaton.

## NOTES

1. This essay was partially prepared while I was a Fellow at the Center for Advanced Study in the Behavioral Sciences. Earlier versions were presented to the Second International Symposium on Informal Logic, University of Windsor, Windsor, Ontario, June 22, 1983, and to a colloquium at Sacramento State University, October 27, 1983. I appreciate helpful suggestions from Peter Gray Whiteley, Robert Linn, William Rapaport, and Andrea Schnall, and am grateful for financial support provided by the Spencer Foundation.
2. One of these is aimed at undergraduate and graduate students (Cornell Level Z); one at secondary and college (Watson-Glaser); and three at grade four through college (Cornell Level Z, New Jersey, and Ross). The Ross test seems to emphasize critical thinking less than the others, as I shall later suggest, but realize that this judgment is based on my conception of critical thinking.
3. Edward Glaser, who has seen these comments and recommendation, was kind enough to provide me with his reaction to them:

"My reaction to your specific questions regarding our scoring key for items 65 and 67 on the Evaluation of Arguments subtest is:

#65. Accepting the argument as true for the purpose of this test, if the actions of a strong labor party would "cause sustained large-scale unemployment," that would be a disastrous consequence for all citizens adversely affected and for our democratic form of government in general. It would seem that only "good" Marxists or other types of revolutionaries who wanted to bring down our form of government and supplant it with their form of dictatorship would consider #65 to be a weak argument in relation to the question posed.

#67. The fact that "labor unions have called strikes in a number of important industries" is **weak** because no information is given about the employees' (or union's) grievances, why they chose to withhold their labor (strike) in those instances, whether their actions were peaceful and legal during the strike period, what results or consequences followed their strike action, etc. A strike **in and of itself** is not **necessarily** "bad" in given instances; the net balance of consequences might be "good" for the country as a whole, for the strikers, and even for the owners of the plants in a given industry over the long run.

If we were rewriting-revising #67, however, I would recommend that the wording be changed to a number of **employer sites** (or companies) rather than 'important industries.'

I agree with you that selecting any particular value position is 'bound to be in conflict with a number of (other possible) value positions.' In a test that explicitly accepts (starts from) the values expressed in our Constitution, Bill of Rights and Declaration of Independence, we do espouse free speech, etc., but what is judged 'good' or 'bad,' or 'strong' or 'weak' arguments with reference to a given issue should (as I see it) be judged from the Judeo-Christian and democratic value orientation underlying our society."

## REFERENCES


Cronbach, Lee J. Test validation. In R.L. Thorndike (Ed.), **Educational measurement** (2nd ed.). Washington, D.C.: American Council on Education, 1971.

Ennis, Robert H. The responsibility of a cause. In B. Crittenden (Ed.), **Philosophy of education 1973.** Edwardsville, IL: Philosophy of Education Society, 1973.

Ennis, Robert H. Identifying implicit assumptions. **Synthese,** 1982, **51,** 61-86.

Ennis, Robert H. & Millman, Jason. **Cornell critical thinking test, level X.** Champaign, IL: Illinois Thinking Project, 1982a.

Ennis, Robert H. & Millman, Jason. **Cornell critical thinking test, level Z.** Champaign, IL: Illinois Thinking Project, 1982b.

Ennis, Robert H. & Weir, Eric. **The Ennis-Weir critical-thinking essay test.** Champaign, IL: Illinois Thinking Project, 1983.

Hertzka, Alfred E. & Guilford, J.P. **Logical reasoning.** Orange, CA: Sheridan Psychological Services, Inc., 1955.

McPeck, John E. **Critical thinking and education.** New York: St. Martin's Press, 1981.

Norris, Stephen P. **A pitfall in the construct validation of ability tests.** Unpublished doctoral dissertation, University of Illinois, U.C., 1981.

Ross, John D. & Ross, Catherine M. **Ross test of higher cognitive processes.** Novato, CA: Academic Therapy Publications, 1976.

Shipman, Virginia. **New Jersey test of reasoning skills.** Upper Montclair, N.J.: Institute for the Advancement of Philosophy for Children, 1983.

Tomko, Thomas N. **The logic of criterion-reference testing.** Unpublished doctoral dissertation, University of Illinois, U.C., 1981.

Watson, Goodwin & Glaser, Edward M. **Watson-Glaser critical thinking appraisal.** New York: The Psychological Corporation, 1980. ●


**Robert H. Ennis, Center for Advanced Study in the Behavioral Sciences, 202 Junipera Serra Blvd., Standford, CA 94305.**