DOCUMENT RESUME

ED 306 225 TM 012 917

AUTHOR Facione, Peter A.
TITLE Assessing Inference Skills.
PUB DATE 20 Feb 89
NOTE 30p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Cognitive Measurement; Construct Validity; Content Validity; *Critical Thinking; Elementary Secondary Education; Evaluation Methods; *Inferences; Knowledge Level; *Multiple Choice Tests; *Standardized Tests; Test Construction; *Testing Problems; Test Validity
IDENTIFIERS Jargon; Technical Language

ABSTRACT

Four major problem areas inhibit the standardized assessment of critical thinking (CT): (1) content validity; (2) construct validity; (3) technical jargon; and (4) background knowledge. Practical examples of framing multiple-choice items for assessment are suggested. In the area of content validity, new agreement about the definition of CT now allows it to be seen as a combination of personality traits, cognitive affects, and cognitive skills. Construct validity is a more troublesome problem because it presumes that one can write questions focusing on the process of thinking as distinct from other factors. Avoiding jargon is a necessity in framing multiple-choice questions. The background knowledge problem can only be addressed by trying to write CT items that presume only the most universal social and human experiences. Specific multiple-choice items that might resolve these difficulties also serve as paradigm frames for targeting three core CT abilities--analyzing, drawing, and evaluating inferences. Every aspect of CT may not be amenable to multiple-choice testing, but the question of whether or not multiple-choice assessment tools framed as suggested above might be suitable is an empirical question that can be addressed by research. A 64-item list of references is provided. (SLD)

# Assessing Inference Skills

Peter A. Facione

Feb. 20, 1989

California State University, Fullerton

## ABSTRACT

This paper discusses four major problems inhibiting standardized CT assessment: The problems content validity, construct validity, technical jargon, and background knowledge. In the process of showing how these problems might be resolved at the practical level, examples of specific MCI's are used. These examples also serve as paradigm question frames for targeting three core critical thinking abilities: analyzing inferences, drawing inferences, evaluating inferences. Example MCI's which require meta-cognition are also included. By way of these example question frames the paper aims to move the philosophical debate over standardized CT assessment toward an empirical resolution.

# Assessing Inference Skills

c. Peter A. Facione

Feb. 20, 1989

California State University, Fullerton

## I: The CT Movement and the Question of Testing

From New Jersey to California, and from Newfoundland to Florida, the leaders of the movement are urging major changes in how we teach and what we test. At all levels our curriculum, our pedagogy and our assessment strategies must form a unified, coordinated emphasis on those trans-disciplinary cognitive dispositions and abilities necessary in this era of information explosion, (Ennis, 1981; Gardner, 1983; Arons, 1983; Sternberg, 1985: Beyer 1985; Quellmalz, 1985; Costa 1985; Ruggiero, 1988; Paul, 1988 (a) and (b)). After decades of relative neglect, throughout the eighties saw a growing accord that the heart of education lies exactly were traditional advocates of a liberal education always said it was -- in the processes of learning and thinking rather than in the accumulation of disjointed skills and senescent information.

But can we validly and reliably assess critical thinking in a standardized format? What might good multiple choice items (MCI's) targeting CT look like? Are labor intensive essay tests the only way to "really" assess CT?

Complex questions like these challenges not just the professor or teacher seeking to introduce critical thinking (CT) goals into his or her classroom, but central offices, boards of education, the educational

testing and publishing industries alike. Since CT is about how students think rather than the answers they produce, sensible assessment strategies must be devised which target how students reason, not what information they have learned. And given the vital importance of teaching thinking, it is little wonder that a great deal of research is being devoted to the topic of CT assessment, (Facione, 1988, Ennis, 1987; Follman, 1987; Pecorino, 1987; Stewart, 1987; Cierzniak, 1985.)

If the heart of CT is process not product, are such widely used standardized tests as the Stanford Achievement batteries and the California Test of Basic Skills sensitive to variations in student's cognitive skills. The research to date suggests not, (Marzano and Jesse, 1987.) Marzano and Costa (1988) report that general cognitive operations required to answer the questions had very little to do with student achievement on those tests. Some skeptics, seeing these results, might be tempted to argue that such tests, because they rely on the multiple-choice format, should never be expected to measure anything but factual or declarative information.

But standardized tests focusing more directly on analytical, logical, or critical thinking skills do exist. The Educational Testing Service boasts a "logical reasoning" section on the LSAT, an "analytical" section on the GRE, "subject-matter based critical thinking questions" on the Advanced Placement Test, "higher order thinking and laboratory-based questions" on the National Assessment of Educational Progress, and items in the in-basket portion of the Foreign Service Test which call for cognitive operations, (Tucker, 1988.) For use at the senior high school level Stephen Norris and R. King, through the Institute for Educational

Research and Development at Memorial University of Newfoundland, developed the *Test on Appraising Observations* 1983. The *Ninth Mental Measurements Yearbook*, lists others.[1]

However, as one soon discovers from reading the reviews, many of these lack either validation or are applicable only in narrow, specialized contexts. What practical advice and workable strategies might the classroom teacher or district assessment director use in writing a standardized CT assessment tool? This paper responds to that question. By way of examples, the paper shows how to frame a variety of CT questions which target core CT abilities in analyzing, evaluating and drawing inferences. Each of these core cognitive abilities is carefully defined as well. Four problems of special note are also defined and discussed -- in the next section content and construct validity, then the Jargon Problem and later the Background Knowledge Problem. Whether or not CT in the context of any given level of education can be assessed adequately using a standardized MC format should not be a theoretical problem for armchair analysis, but an empirical problem for educational research. By giving practical examples of how MCI's can be framed, this paper will advance the issue in that direction.

## II: Validity Problems and CT Assessment

In spite of the familiar criticisms and concerns regarding multiple choice testing, the case in favor of multiple choice (MC) testing of crucial aspects of CT remains solid, particularly if MC instrumentation is conceived of in an appropriately restricted way. How might that be?

Namely, as *MC testing is but one efficient and practical way of gathering reliable evidence regarding large numbers of persons from which evidence one might, with justifiable confidence, draw inferences regarding the relative abilities of those persons on some, but not necessarily all, important dimensions of critical thinking,* (Norris and Ennis, 1987; Sternberg, 1987; Kearney, 1986; Facione, 1984). This is not to say, however, all the important theoretical questions regarding MC CT testing have been put to bed.

The first key issue of content validity, for example, is still a major concern. Exactly what is it that we are talking about when we proclaim to be teaching and assessing "critical thinking?" Is CT discipline-neutral or discipline-specific? Is CT an ability or set of skills only, or does CT also require that a person display certain attitudes or dispositions?[2] The debate among CT experts, while still not resolved, seems to have hit on some vital area of accord. For one, CT is best viewed as a combination of dispositions (personality traits, cognitive affects) and abilities (cognitive skills). Both elements can are evident in the most influential definitions of CT. Second, there is accord on some crucial descriptors, for example that it involves making reasoned judgments using relevant criteria, it analytical, it is evaluative of the product of thought, it is self-consciously meta-cognitive. (Dewey, 1909; Ennis, 1987; Glaser, 1941; Lipman, 1988 (a); McPeck, 1981; Paul, 1988 (a)). And third, influential theoreticians like Ennis, Lipman and Paul amply supply us with rich descriptions of CT, with strategies for introducing CT into the curriculum, and with many seperate aspects of CT which might be assessed, (Ennis, 1987; Paul 1988 (b); Lipman 1988 (b)). So, finding a padagogically useful conceptualization

of CT in which to ground one's CT assessment is not the problem it once was, even if consensus and accord still elude the experts.

A second, and more troublesome issue in CT assessment is construct validity. In terms of MC testing, the question can be posed this way: How can we be sure that selecting the keyed response indicates a correct application of CT skills and selecting any of the distractor responses indicates an incorrect application? Might it be that keyed answers are selected for wrong reasons and distractors are selected in spite of good thinking? And if so, why and how can the source of the invalidity be found and corrected?

There are two different ways of addressing the problem of construct validity. The traditional way is for experts to analyze test items and judge what cognitive operations achieving a correct answer on that item would require. On a CT test, of course, this becomes a matter of hypothesizing what students should have been thinking to select keyed answers. But these hypotheses might be mistaken. Moreover, recent research in cognitive development suggests one key difference between experts and novices in any given field is how they approach problems in that field. If that is the case, then relying on experts to say how novices (students) ought to think through a given CT test item may not be a sufficient guarantee of construct validity. Fortunately new strategies for judging construct validity are being developed by researchers like Steven Norris. Instead of depending on *a priori* suppositions, Norris' approach is to check construct validity *a posteriori*, by direct interactions with those subjects for whom the test is targeted. Subjects are interviewed while in the process of answering pilot items, and they

are asked to describe what they are thinking as they consider and select their answers. In this way, a more direct kind of evidence regarding why subjects make the choices they make can be gathered, (Norris 1988.)

The challenge of framing clean MC CT items relates to both validity problems: (a) to content validity for it presumes that one has a clear idea about which cognitive skills are included in CT and which are not, and to (b) construct validity because it presumes that one can write questions which focus on the process of thinking as distinct from other factors -- such as the content thought about or the vocabulary used -- which might lead students to select the right answer for the wrong reason or select the wrong answer for the right reason. Whereas theoretical arguments about the possibility of meeting this challenge might be mounted, this paper aims at a more practical response. By suggesting possible examples of MC CT questions, the issue is transformed into an empirical, not a philosophical, question.

All of the currently prominent conceptual analyses of CT maintain that, whatever else it might include, CT is centrally includes cognitive skills analyzing inferences, drawing inferences and evaluating inferences.[3] Granted that there may also be a concomitant set of intellectual attitudes and dispositions associated with CT, and granted CT plays crucial and complex roles in a wide variety of different human pursuits where discipline-specific background knowledge is crucial to its successful application, the cognitive abilities of analyzing, evaluating, and drawing inferences cut across subject fields, educational levels, and specific applications. Thus, for CT assessment, much of the problem of standardization is solved if one can develop MCI's which validly and

reliably target precisely those core CT cognitive skills and sub-skills.

## III: The Jargon Problem and Assessing CT

The artist who has all the dispositions and skills to do beautiful work, but who lacks the abilities to describe how she or he achieves such success is not unlike the person who has all the cognitive attitudes and skills to be a good thinker but lacks a knowledge of the technical vocabulary to describe how she or he achieves CT success. We would not say that the artist is any less an artist for being poor at describing the artistic temperament or rightly naming artistic skills. Why then would we want to make CT assessment depend on being able to talk about CT the way cognitive psychologists, logicians or philosophers talk about it?

Using this analogy we can begin building our CT assessment tools by ruling out questions which target CT vocabulary or the specialized academic language used in talking about logic or CT.[4] Thus the following would *not* be an acceptable CT question: "Which of the following is a valid rule of deductive inference: *A= Disjunctive Syllogism, B= Generalization, C= Circular Reasoning, D= Equivocation." Nor would the following: "When a person argues that his view must be correct simply because no one has brought up a good reason why it is wrong, the person is said to be committing the fallacy of: A= Attacking the person, B= False Cause, C= Begging the Question, *D= Appeal to Ignorance."

However, it might be objected, even CT has some vocabulary! It may

not be vital to name rules of inference or species of fallacies, but there have to be ways to talk to students about "evidence," "premises," "conclusions," "arguments," "credibility," "validity," "deduction," "induction," and the like. So why not use questions like the two above? Would it not be far easier to write questions using technical vocabulary? Of course, it would be. And perhaps in the context of a specific CT curricular program and for specific evaluative purposes that might be reasonable to do.

However, there are at least three independent reasons to resist mightily the temptation to use technical vocabulary on CT tests --- particularly on CT tests which are intended, like general aptitude or ability tests, to be used outside the context of any given course or curricular program. First, some CT vocabulary already exists in our language. If teaching CT leads to more precision in its use, that would be a desirable side benefit. But creating a technical vocabulary of CT (which really serves little purpose beside to distinguish the initiated from the uninitiated) has the devastating effect of transforming CT just another "school subject" for students "to study." This can only harm efforts to infuse the curriculum with CT.

Second, people can be very good at CT without having mastered the use of these words in their technical senses. If we are aiming at testing "native" CT ability, we should strive to do so without making the demonstration of that ability vocabulary-dependent. By analogy, it is no more necessary to ask questions about technical logic or CT vocabulary on a CT test than it is to ask questions about technical psychology or education vocabulary on an intelligence test or a reading test. As the

examples given later show, one need not use words like "valid" or "justified" in their technical senses to ask questions which target the quality of an inferences. As it turns out, knowing about CT and being good at CT are different things.

Third, if the particular CT skill or sub-skill one aims to assess can only be focused on by using a word which may be misunderstood by a significant number of the persons for whom the test is being designed, then it may be possible to define that word in the context of the question, thus avoiding if possible having respondents miss the item due to vocabulary, rather than thinking process, deficiencies. Consider the following example. It was given to 108 college level general education students at the end of the fourth week of a 16 week semester CT course. It targets the analytical CT skills of (a) distinguishing arguments from non-arguments and, (b) given an argument, identifying its conclusion. In the context of this curricular program it was reasonable to use the word "argument" in the test item.

8. 'To judge the morality of an action one need only look at its consequences. Some actions have beneficial consequences, others do not. Killing an innocent person might be a great benefit to society. So, killing an innocent person can be morally correct.'
This passage is:
A= Not an argument.
B= An argument, the first sentence is its conclusion.
C= An argument. the second sentence is its conclusion.
*D= An argument. the fourth sentence is its conclusion.
[Percentages correct: Total = 82, Top 27% = 90, Bottom 27% = 69.]

But if the intention had been to do a CT pretest, or if it were suspected that a significant number of those persons for whom this test was targeted might be confused by the use of words like "conclusion" or "argument," then the question could have been framed this way:

8'. Consider the following passage: "(1) To judge the morality of an action one need only look at its consequences. (2) Some actions have beneficial consequences, others do not. (3) Killing an innocent person might be a great benefit to society. (4) So, killing an innocent person can be morally correct." Which sentence, if any, does the author present as his main contention or claim which he supports by using other sentences in the group?
A= None B= (1), C= (2), D= (3), *E= (4).

## IV: MCI's Targeting Analytical CT Skills

The critical thinking skill of *analyzing* involves identifying the inferential relationships between statements, descriptions or representations which express reasoned judgments, beliefs or opinions. Analyzing includes two sub-skills: *locating arguments* and *parsing arguments*. Given a set of statements, descriptions or representations determine whether this set expresses or was intended to express a reason or reasons in support of some claim, opinion, or point of view. Given the expression of a reason or reasons in support of some claim, opinion or point of view, identify: (a) an argument's intended conclusion, (b) the premises or reasons the author advanced intending to support that conclusion or to back-up other premises the author uses in support of the intended conclusion, (c) additional unexpressed elements of that reasoning, such as intermediary conclusions or unstated assumptions, and (d) for exclusion, any items contained in the body of expression being analyzed which are not intended to be taken as inferentially crucial to the reasoning being expressed.

Like the example questions above, one straightforward MCI question frame for assessing analyzing begins by giving a passage and asking students for an interpretation of that passage or for the inference role

played by any given sentence in that passage. Or, suppose the focus is on identifying an intended but unstated premise or conclusion -- one the author has omitted believing it to be too obvious to bother mentioning in a given context. The following targets such an unspoken assumption.

13. "Many specialized departments have been developed recently by the Breem Corporation. This proves that Breem Corp. is very interested in more sophisticated approaches to reaching the marketplace." This passage is best described as:
A= Missing the conclusion: "Breem Corp. is now doing a better job of reaching the marketplace."
*B= Missing the premise: "These new departments are working on sophisticated approaches to reaching the marketplace."
C= Missing the premise: "Breem's stockholders thought that new approaches to reaching the marketplace were necessary."
D= Missing the premise: "Breem Corp. was not reaching the marketplace before these new departments were developed."
E= Not an argument.
[Percentages correct: Total = 76, Top 27% = 97, Bottom 27% = 45.]

Question frames which focus on skill-integration, even at the analytical level, are highly desirable, particularly since CT skills must be well integrated if they are to be anything but dysfunctional or counterproductive. The following are examples of questions aimed at seeing how well the same group of college students in the fourth week of their 16 week course, could work through something like the following progression of sub-skills: (1) distinguish passages which contain arguments from those that do not, then, (2) if a passage contains an argument, identify the conclusion of that argument, then (3) distinguish among argument passages those which offer a single reason for the conclusion from those which offer multiple independent reasons for the conclusion, then (4a) within single-reason passages, distinguish various interrelated premises, or (4b) within multiple reason passages, distinguish separate reasons, and finally (5a) supply obvious but unstated conclusion or premises or (5b) identify the inference role

played by a particular statement in the passage. Here are two example test items:

17. "Come on. There's nothing wrong with cheating. Look around! Everybody does it. And besides, what harm can come from one miserable freshman cheating a little in a general education course. I mean, it isn't like the fate of the world depends on what grade I get in Introduction to Philosophy." This passage is:
A= Not an argument.
B= An argument giving only one reason.
C= An argument with the conclusion: 'The fate of the world does not depend on what grade I get in Introduction to Philosophy."
D= An argument with the missing conclusion: "I am not majoring in Philosophy."
*E= None of the above.
[Percentages correct: Total = 62, Top 27% = 83, Bottom 27% = 38.]

25. "It is detrimental to science education to teach religious ideas mislabeled as science. This is so because it misleads our youth about the nature of scientific inquiry. Scientific inquiry does not permit one to believe an hypothesis which has been proven false. Creationism based on a very literal interpretation of the Bible is such a false hypothesis. So, belief in creationism is not scientific. Which means that it would be misleading to our youth to present creationism as the product of scientific inquiry. Besides, teaching religious ideas mislabeled as science also strips our citizens of the power to distinguish between the phenomena of nature and the articles of faith." This passage is:
A= An argument with its main claim being its last sentence.
*B= An argument which provides exactly two reasons for its main claim.
C= An argument which provides exactly one reason for its main claim.
D= An argument with the main claim being 'Teaching religious ideas mislabeled as science misleads our youth about the nature of scientific inquiry."
E= An argument based on the assumption that evolution is true.
[Percentages correct: Total = 28, Top 27% = 52, Bottom 27% = 07.]

## V: MCI's Targeting Skills in Evaluating Inferences

The critical thinking skill of *evaluating* involves assessing the credibility of statements, descriptions, or representations, and assessing the strength of the inferential relationships between claims

and the reason or reasons advanced in their support. Evaluating includes two sub-skills: ***verifying claims*** and ***assessing logical strength.*** Verifying claims involves assessing the degree of confidence to place in a given statement, description, or representation. Assessing logical strength involves determining the nature and quality of inferential relationships by judging whether the assumed truth of the premises of a given argument justify accepting as true, or very probably true, the conclusion of that argument.

Differences in background knowledge and cultural presumptions are always complicating factor whenever inference evaluation skills are the target of assessment. Why? Because CT does not occur in an intellectual and human vacuum. It must be *about* something. Yet, can any classroom instructor claim his or her students all share the same intellectual traditions, academic background information, and cultural presumptions? No. The *Background Knowledge Problem* in a pluralistic culture can be addressed only by trying to write CT items which presume only the most universal social and human experiences (within the life experiences of one's students) and which also supply sufficient information in the question stem to reasonably assure that the intended respondents have sufficient information to correctly evaluate the inferences being critically examined.

Citing the *Background Knowledge Problem* is not meant as a criticism of the educational system. With the explosion of knowledge in so many fields, it is clearly wrong-headed to conceive of of education as fact-loading. It is equally impractical to think that the goal is to equip the entire population with a unified body of academic background

information. Education is not a hammer to enforce some misguided goal such as cultural homogeneity. *An argument for infusing CT into the curriculum is that learning how to think, not what to believe, is a main goal of education.* How counterproductive it would be to demand that to show well on a CT test students would have to know what to believe, not how to think!

Even if one finds a MCI topic where background information is shared, different cultural assumption or interpretations of those assumptions might still lead people to correctly (logically) infer different conclusions. After all, baseball is baseball and fair is fair! Right? So, consider the following question: "When a stud hitter comes to the dish it would be fair for blue to: A= Expand the strike zone, B= Squeeze it, C= Be sure to leave the strike zone unchanged." Even if students know baseball and baseball slang, they still cannot be sure of the right answer unless they also know where the baseball game is being played and what cultural interpretations are put on "fairness" in that place. That is why there is no keyed response. The "right" answer in the USA is C, but in Japan it is A. (In Japanese baseball, fairness demands that the strike zone for good hitters be made larger to balance out their superior skill and make things fairer in the pitcher/hitter competition.) The only reliable way to check on one's assumptions is to verify *a posteriori* the construct validity of each CT test item.

One might be tempted to argue that the *Jargon Problem* and the *Background Knowledge Problem* show why MC testing is an inferior mode of CT assessment. But nothing gives essay tests any greater immunity from these problems. Little is gained by forgoing MC testing simply because

of these two concerns. Further complicating the essay test strategy for assessing CT skills are the notorious difficulties of separating specific skills being tested, test reliability, the imprecision of test results, and the impracticality of labor intensive essay testing.

The *Watson-Glaser CT Appraisal* uses a three part question frame to test inference-evaluation. The first part presents information, the second a proposed inference, and the third part invites responses to the a question such as "Given the information above, is the inference drawn: A= true, B= probably true, C= probably false, D=false, or E= unknown." An advantage of this frame is that the answer selections can be held constant through a large number of items, thus permitting greater familiarity and fewer instrumentation difficulties. A difficulty, however, is that it permits no comparisons between alternative possible inferences which plausibly might be drawn from the same body of information. The following example, taken from the college level CT exam cited earlier, reduces the Watson-Glaser three part question frame to two parts.

35. Suppose you have a standard deck of 52 playing cards. The deck contains exactly four kings, four queens and four jacks. For our purposes we will call these twelve cards the only "face-cards" in that deck. Suppose you shuffle the deck of 52 cards and are about to randomly draw one card. Given this set up, what can be logically inferred?

A= That you will necessarily draw a face card.
B= That you will probably draw a face card.
C= That you cannot possibly draw a face card.
*D= That you probably will not draw a face card.
E= Nothing can be logically inferred about your drawing a face card.

[Percentages correct: Total = 71, Top 27% = 93, Bottom 27% = 52.]

In addition to global inference evaluation, specific evaluation sub-skills can be targeted using the MC format. Here, for example, is a

question from an examination given in the same college CT course mentioned above. 98 students took this exam, which was given at the end of the eighth week of the semester. Its questions target identifying and classifying fallacies by name, something which is reasonable to expect of students since they had been taught the vocabulary, but something which should be avoided because it introduces the problem that a student might be able to judge why an inference is faulty but miss the item because he does not know or remember the right fallacy name.

23. In this passage consider Christopher's reasoning: "In the half light of pre-dawn little Christopher J. sat quietly with his nose pressed against the cool glass of his bedroom window. He wanted very much for it to be morning so he could go outside and play baseball. Concentrating very hard, he wished and wished for the sun to appear. And as he wished the sky began to brighten. He kept wishing. And, sure enough, the sun moved right up over the horizon and into the morning sky. He was proud of himself. Christopher thought about what had happened and decided he could make any cold and lonely night turn into a bright and happy summer day, if he wanted."

A= Fallacy of playing with words or playing with numbers
*B= Fallacy of false cause, false dilemma, or gambler's fallacy
C= Fallacy of composition, division, or distribution
D= Fallacy of the straw man, or fallacy of no logical progress
[Percentages correct: Total = 81, Top 27% = 88, Bottom 27% = 58.]

However, winning the "Name That Fallacy" is hardly the main business of inference evaluation. Items like those above would be much improved if the names of specific fallacies were replaced with descriptions of the kinds of mistakes the names denote. Nor should CT inference evaluation be confined to the short staccato bursts characterized by discrete MCI's like those above. More sustained and complex contexts must be provided, particularly in assessing college or adult level CT ability. Here, for example, is a series of questions calling for a broad range of inference evaluation sub-skills. This question frame begins by granting that the inference under examination is faulty and goes on to ask why. The frame

is useful for focusing on judgments regarding logical strength, reliance on assumptions, and the relevance of information.

The following sample questions were used on the final examination for the same college level general education CT course. (N = 104) In addition to asking students to evaluate an inference as good or bad, they focus the student's attention on the reasons why the inference is good or bad. There are better and worse defenses for that judgment. By asking for good reasons why a good inference is good, the question also calls for the activation of another crucial CT ability: *meta-cognition.* In effect, the frame below demands that students think about thinking.

Consider the faulty inference in the following fictional case: "A study of 3400 autos currently in use by six auto rental companies, in Arizona, New Mexico and, Texas revealed that 30% of these autos were not able to meet the 1987 US Government Standards for air pollution control. All the cars studied were built in the USA in 1987. According to eighteen administrators interviewed (three working at each rental company), all companies have the policy that pollution control equipment is to be checked, and where needed, repaired every 10,000 miles. In mileage, the 3400 cars studied ranged from a low of 18,000 to a high of 28,000. Based on this data, the researcher claimed that 30% of all 1987 model cars operated in the United States would fail to meet the same government standards once they had been driven 23,000 miles, even if their pollution control equipment had been checked and repaired, if needed, every 10,000 miles."

17. One rental agency executive said, "The inference from these data to the claim being made is faulty because a significant number of foreign built cars are operated in the US and these cars might have superior engineering." If true, would the executive's reason be a good reason or a bad one, and why?

A= Bad reason, engineering does not relate to pollution control.

*B= Good reason, the study drew an inference about 1987 foreign cars operated in the US, but didn't study any.

C= Bad reason, the researcher did not propose any conclusion about cars built outside the USA.

D= Bad reason, the data collected only relates to US built cars, so talking about foreign cars is irrelevant.

E= Good reason, everyone knows foreign cars have superior pollution control engineering.

[Percentages Correct: Total = 49, Top 27% = 75, Bottom 27% = 11.]

18. An auto insurance agent said, "The inference is faulty because all the cars in the sample were fleet cars, none were privately owned and operated." If true, is the auto insurance agent's reason a good or a bad one, and why?

*A= Good reason, this factor might be relevant and should have been considered.

B= Good reason, fleet cars receive periodic inspections, but privately owned cars do not.

C= Good reason, privately owned cars are less likely to have been tampered with than fleet cars.

D= Good reason, cars driven in Arizona, New Mexico and Texas are driven more recklessly.

E= Bad reason, who owns a car in not relevant to whether or not the engine performs up to government standards.

[Percentages Correct: Total = 47, Top 27% = 61, Bottom 27% = 29.]

21. A newspaper editor from New Mexico said, "The inference is faulty because there is reason to think auto rental companies may never have actually conducted the pollution equipment checks or made the repairs, or they may have falsified the data regarding the regularity of their safety checks and repairs." If true, is the editor's reason a good one or a bad one, and why?

A= Bad reason, the number or regularity of the checks is actually irrelevant.

B= Bad reason, there is no evidence that anybody has lied or has any vested interest in falsifying such things.

C= Bad reason, the regularity of safety checks and repairs was solidly established by the interviews.

*D= Good reason, just because a company has a policy doesn't mean the policy is actually carried out.

E= Good reason, the executives may have been lying about what policies their companies actually had.

[Percentages Correct: Total = 64, Top 27% = 86, Bottom 27% = 50.]

## VI: MCI Frames Targeting Skills in Drawing Inferences

The critical thinking skill of *inferring* involves securing the elements needed to make inferences and determining the inferential relationships between or flowing from statements, descriptions or representations. Among the sub-skills of inferring are ***querying, conjecturing, and drawing conclusions.*** Querying involves recognizing the need for evidence or information and formulating a strategy for seeking and gathering that evidence or information. Conjecturing involves formulating alternatives, developing hypotheses, and postulating

suppositions. Drawing conclusions involves deducing or inducing the logical consequences which are implied, entailed, warranted, supported, by of a given set of statements, descriptions, or representations.

One question frame to target this essential CT skill is a modification of the three part Watson-Glaser frame described above. The first part supplies the information, the second offers a series of four plausible alternative inferences which might be drawn, and the third asks "Assuming that the information provided is true, which of the above claims could not possibly be false?" or "... is very probably, but not necessarily, true?"

In an essay test or short answer format it seems plausible to focus on drawing conclusions by providing a case study followed by a set of interrogations inviting students do draw inferences from the information and principles presented.[5] To induce respondents to think proactively using the MC format, one can use a question frame which begins by inviting the respondents to initiate inferences. This can be accomplished by modifying the suggestion in the paragraph above, transposing the query and the body of information. The MCI would begin with a question, for example, "What is the most reasonable, non-fallacious inference to draw given the following...?" This way the subject is invited to anticipate the answer choices and attempt to draw the proper inference before being prompted by reading the right and wrong choices. Certainly, having any prompts at all makes drawing inferences in the MC context different, if not easier, than if no prompts, right or wrong, were provided. It remains to be demonstrated, however, that this apparent shortcoming is severe enough to render MC tests inappropriate

for gathering sufficient evidence to make good judgments regarding the respondents inferencing abilities. And, at each educational level from elementary to post-graduate, this question is no longer philosophical, but empirical. The burden of proof, it would appear, falls equally on those who would defend the MC mode or the essay mode as superior.

The sample inference drawing MCI below were administered during week twelve of the same college general education CT course. (N = 103)

2. What is the most reasonable, non-fallacious inference to draw from the following: "Song writers are very rich people. If a person is very rich, he must devote a great deal of time to managing his money. But if he has that much money, he would benefit from a degree in Business. So..."

A) poor song writers don't need Business degrees.
B) song writers are rich people.
C) song writers with money must devote time to managing it.
D) any who benefit from a degree in Business, are rich.
*E) song writers would benefit from a degree in Business.

[Percentages Correct: Total = 83, Top 27% = 96, Bottom 27% = 71.]

Drawing inferences is a complicated business and in different contexts it can involve different sub-skills. Preparing to defend one's opinions by anticipating objections and developing responses is one such context. Conducting a strategically delicate cross-examination is another. Engaging in scientific research is still another. One advantage of MC testing is that it permits focusing more directly on some of these sub-skills. For example, in the context of drawing inferences regarding empirical phenomena, some of the sub-skills include being able to (a) identify issues requiring the application of specific empirical research techniques informed by the appropriate background knowledge, (b) define the nature of the background knowledge needed to decide a given issue, (c) generate plausible hypotheses regarding a given issue, (d) conceive of procedures for testing a given hypothesis relative to a given

issue, and (e) determine which competing hypothesis would have to be ruled out to strengthen one's confidence in a test hypothesis. Targeting some of these kinds of inference drawing sub-skills, as well as some of the concepts needed to discuss these skills, the following MCI's were given as part the earlier mentioned final examination in a college level CT course.

Consider the following fictional research report: "Research at the Experimental Nursery School on the campus of State University, showed that four-year-old children who attended the Child Care Center all day for 9 months averaged 58 points on a standardized test of kindergarten readiness. The research showed also that those four-year-olds who attended only in the morning for 9 months averaged 52, and those four-year-olds who attended afternoons only for 9 months averaged 51. A second study of four-year-olds who attended Holy Church Nursery School all day for 9 months showed these children averaged 54 on the same kindergarten readiness test. A third study of four-year-olds who attended neither nursery school nor day care centers showed an average score of 32 on the same test. The difference between 32 and the other scores was found to be statistically significant at the .05 level of confidence."

4. To scientifically disconfirm that there is no correlation between attending pre-school and kindergarten readiness one would have to do which of the following?

A= Find that 95% of all four-year-olds were kindergarten-ready.

B= Find a child who was kindergarten-ready but who did not attend any nursery school or day care center.

*C= Find that there is less than 5% chance that the connection between attending pre-school and kindergarten readiness is random.

D= Find that attending pre-school is causally related to earning good grades in high school.

E= There is no way to scientifically disconfirm it.

[Percentages Correct: Total = 76, Top 27% = 86, Bottom 27% = 64.]

5. Assume a researcher advanced the hypothesis that, given the data above, "full time attendance in an organized pre-school program increased a child's readiness for kindergarten by about 40%." Which of the following alternative hypotheses would have to be <u>ruled out</u> in a well-designed experiment?

1. The children studied were the children of affluent, professional people who could afford nursery school tuition and so, these children could be expected to be better prepared for kindergarten than the children of average or low income parents.

2. The "experimental" nature of the State University Nursery School biased the outcome as compared to more standard pre-school experiences.

3. Since none of the children studied were less than four years old, the study proves only that a pre-school experience benefits children who are four-year-olds.

4. Parents of slower children do not send them to organized pre-schools, so the population is pre-selected for higher kindergarten readiness, but being in pre-school does not really do anything for the children.

Choices: A= 1, 2 and 3.
B= 1, 3 and 4.
C= 2, 3 and 4.
D= 2 and 3.
*E= 1, 2, 3, and 4.
[Percentages Correct: Total = 38, Top 27% = 64, Bottom 27% = 18.]

## VII: Conclusion

This paper has provided examples of question frames designed to focus on three core CT skills areas: analyzing inferences, drawing inferences, evaluating inferences, and, to some extent, metacognition. These samples are intended to strengthened the case for standardized MC testing of certain vital aspects of CT. Granted, not every aspect of CT is may be ameanable to MC testing. Assessing CT dispostions, attitudes or cognitive traits, poses significantly greater challenges. Likewise, no completely suitable standardized CT instrument for use at the particular educational level in which one is interested might now exist. But whether or not MC assessment tools using MCIs framed as suggested above might be suitible, is an empirical question for educational research. In that respect, the issue has advanced from the philoophical and theoretical to the practical. That is the goal of this paper.[6]

---

ACKNOWLEDGMENTS

This paper evolved most immediately from a presentations at Sacramento State University entitled "MC Strategies for Testing CT," "Thirty Great Ways to Mess Up a CT Test," and "Defining CT for Practical Purposes." The views

expressed also have been shaped by over a year of coordinating interactive Delphi research on CT assessment as well as nearly twenty years of classroom experimentation with MC CT question frames. I am particularly grateful to Donald Scherer of Bowling Green State University with whom I worked on these issues for many years. I must also thank the students at Bowling Green and at California State University Fullerton for responding to these test items and particularly for pointing out the many flaws, ambiguities and infelicities in how they might have been framed. I am grateful also to Berry K. Beyer of George Mason University for his helpful comments on an earlier draft.

---

## FOOTNOTES

[1] Among the listings in the Ninth Mental Measurements Yearbook will be found: #269 "Cornell Critical Thinking Test," (Ennis, Millman, Tomko), 1961-1983, Illinois Thinking Project, University of Illinois, Urbana. Reviewed in *Educational and Psychological Measurements* 1983 Vol. 43., pp. 1187-1197, by Modjeski and Micheal. #390 "Ennis-Weir Argumentation Test, Level X: An Essay Test of Rational Thinking Ability," (Robert Ennis and Eric Weir) 1982, Illinois Thinking Project, University of Illinois, Urbana. Reviewed by Herbert Rudman, Michigan State, in NMMY. #391 "Ennis-Weir Critical Thinking Essay Test: An Instrument for Testing/Teaching," (Robert Ennis and Eric Weir) 1983. #1347 "Watson-Glaser Critical Thinking Appraisal" 1942-80. Described and reviewed by two persons in the NMMY, many citations of other research regarding this instrument. #751 "New Jersey Test of Reasoning Skills," 1983, Virginia Shipman, Institute for the Advancement of Philosophy for Children. #1258 "Test of Inquiry Skills" 1979, Australian Council for Educational Research. For junior high grades, this test purports to evaluate a range of research, study and critical thinking skills in the sciences. #1061 "Ross Test of Higher Cognitive Processes," (John Ross and Catherine Ross) 1976-79, Academic Therapy Publications. For grades 4-6, this test includes sub-scores on analogies, deductive reasoning, missing premises, questioning strategies, and relevance of information. #1248 "Test of Cognitive Skills" 1981, McGraw Hill. For grade levels 2-12, this test includes sub-scores on sequencing, analogies, memory, and verbal reasoning. #1269 "Test of Problem Solving" 1984, LinguiSystem Inc. For ages 6-12, this tests a child's thinking and reasoning abilities critical to events of everyday life. It includes sub-scores on explaining inferences, determining causes, negative why questions, etc. #272 "Corrective Reading Mastery Test" 1980, Science Research Associates, Inc. Designed to measure the effectiveness of corrective reading programs, this test includes sub-scores on deductions, classifications, analogies, inductions, statement inference, hypothesis/evidence. #1302 "Deductive Reasoning Test," (J. M. Verster) 1972-73, National Institute for Personnel Research, South Africa. Focuses on syllogistic problems and designed for for candidates for graduate scientists and higher professions. #1010 "PSI Basic Skills Test for Business and Industry" 1981-1982, Psychological Services Inc. Includes sub-scores on problem solving, decision making, reasoning and classifying. #106 "Ball Aptitude Battery" the Ball Foundation. Used to tests persons for occupational placements, this instrument includes sub-scores on inductive reasoning, analytical reasoning, idea fluency, and shape assembly.

[2] We immediately run into problems of content validity. What exactly is

CT? Assuming that skills in drawing inferences and evaluating inferences are main goals does not preclude other main goals. Currently I am coordinating a research project, begun in January of 1988, regarding content validity. This project, sponsored by the American Philosophical Association's Committee on Pre-College Philosophy, is attempting, through the use the Delphi process and a cross-disciplinary panel of sixty North American experts, to come to some accord regarding the core operations in the concept of Critical Thinking. If accord is reached, this could move the issue of content validity much closer to resolution and could provide a clear focus for CT assessment.

[3] This assumption evolved from my earlier papers published in *Liberal Education* and could stand in need of further modification depending on the Delphi results mentioned in note 2.

[4] As obvious as these two points are, in practice we still make these mistakes. How many of us in writing CT test items for use in our own classrooms fall back into straight memory and vocabulary questions? One way to avoid these errors is to ask as one writes the test item, "Can my students answer this without having to know any *special* facts or vocabulary?"

[5] The word "reasonably" is essential here. It is intended to rule out two paradoxical quirks of logical theory. The first is that an inconsistent set of premises logically implies anything at all. The second is that given the rule of inference which sanctions inferring "Either A or B" from "A" an infinite number of irrelevant but logically correct conclusions can be drawn from any one statement.

[6] After reading a draft of this paper a colleague commented, "Well, I'm convinced you can test CT, now my concern is whether CT can be learned! But one of the values of getting some good testing instruments will be just that, to find out if it can be learned."

---

BIBLIOGRAPHY


Arons, Arnold B. "CT and the Baccalaureate Curriculum," *Liberal Education*. v71, n2. Summer 1985.

_____, "Achieving Wider Scientific Literacy," *Daedalus, Journal of the American Academy of Arts and Sciences*, v112, n2, p91-122, Spr. 1983.

Baron, Joan B., and Sternberg, Robert J., *Teaching Thinking Skills: Theory and Practice*. W. H. Freeman Publishing, 1987.

Beck, Ronald A., *A Guide to Criterion-Referneced Test Construction*, Johns Hopkins University Press, 1984.

Beyer, Barry K., "Improving Thinking Skills -- Defining the Problem," *Phi Delta Kappan*, v65, n7, p486-90, March 1984.

_____, "Practical Strategies for the Direct Teaching of Thinking," in *Developing Minds: A Resource Book for Teaching Thinking*, Arthur L. Costa, ed., Association for Supervision and Curriculum Development, Alexandria,

VA, 1985.

_____, "A Suggested Format for Testing Thinking Skills," *Social Science Record*, v24, n1, p3-5, Spr. 1987.

Block, R. A., and Taylor, S. V., "Cognitive Skills: Enhancement and Assessment Issues," Presented to the American Psychological Association, Toronto, Canada, 1984.

Bloomberg, Fran, *et. al. A Pilot Study of Higher-Order Thinking Skills Assessment Techniques in Science and Mathematics -- Part I and Pilot-Tested Tasks -- Part II, Final Report*, National Assessment of Educational Progress, Princeton, NJ, Nov. 1986.

Bransford, J. D., *et al*, "Teaching Thinking: Evaluating Evaluations and Broadening the Data Base," *Educational Leadership*, v44, p67-70, 1986.

Braungart-Bloom, Diane S., "Assessing Higher Thinking Skills through Writing," 10p. Presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA, April 1986.

Carlson, E. R., "Implications of Cognitive Theory and Research for Teaching CT, Presented to the Annual Meeting of the American Psychological Association, Toronto, Canada, 1984.

Cierzniak, Susanne Lipetska, *The Question of CT: An Annotated Bibliography*, 64p. Exit Project, University of Indiana at South Bend, ED 260 069, April 1985. The specific focus of this work is CT in the secondary schools.

Costa, A., ed. *Developing Minds: A Resource Book for Teaching Thinking*, Assoc. for Supervision and Curriculum Development, Alexandria, VA, 1985.

Dewey, John, *How We Think*, 1909.

Ennis, Robert H., "Testing for CT: State of the Art," American Educational Research Assoc., San Francisco, CA, 1968.

_____, "Rational Thinking and Educational Practice," in *Philosophy and Education*, Jonas Soltis (Ed.), Eighteenth Yearbook of the National Society for the Study of Education, Chicago, NSEE, p143-183, 1981.

_____, "Problems in Testing Informal Logic CT Reasoning Ability," *Informal Logic* v6, n1, p3-9, 1984.

_____, "A Logical Basis for Measuring CT Skills." *Educational Leadership*, v43, p44-48, 1985.

_____, "A Taxonomy of CT Dispositions and Abilities," in *Teaching for Thinking*, Joan Baron and Robert Sternberg (Eds.), Freeman, New York, NY, 1987.

_____, "A Bibliography of Testing CT," *CT News*, Center for the Reasoning Arts, CSU Sacramento, v6, n1, Sept.-Oct. 1987.

Ennis, Robert H., and Norris, Stephen P., "CT Testing and Other CT Evaluation:

Status, Issues, and Needs," in *Issues in Evaluation*, Algina, James, (Ed.), Ablex Press, New York, NY, 1988(?).

Facione, Peter A., "Toward a Theory of CT," *Liberal Education*, v70, p253-261, Fall 1984.

_____, "Testing College Level CT," *Liberal Education*, v72, p221-232, Fall 1986.

_____, "Teaching about Fallacies," *Teaching Philosophy*, v10, pp. 211-217, Summer 1987.

_____, "A CT Bibliography with Emphasis on Testing," *CT News* forthcoming, Winer 1989.

_____, "Some Current Questions regarding HOTS Assessment," *CT News* forthcoming, Winter 1989.

Follman, John, "Contemporary CT Bibliography," *CT News*, Center for the Reasoning Arts, CSU Sacramento, v6, n2, Nov.-Dec. 1987.

Glaser, Edward M., *The Development of CT*, 1941.

Glaser, R, "Education and Thinking: The Role of Knowledge," *American Psychologist*, v39, p93-104, 1984.

Kearney, C. Philip, *et al*, " Assessing Higher Order Thinking Skills," ERIC Clearinghouse on Tests, Measurement, and Evaluation, Princeton NJ, p86 April 1986

Lipman, Matthew, "CT: What Can It Be?" *Educational Leadership*, Sept. 1988 (a).

_____, "Characterizing CT," Center for Philosophy for Children, Montclair State College, NJ. 1988 (b)

Marzano, Robert J, and Costa, Arthur L., "Question: Do Standardized Tests Measure General Cognitive Skills? Answer: No," *Educational Leadership*, May 1988, pp 66-71.

Marzano, Robert J, and Jesse, D., "A Study of General Cognitive Operation on Two Achievement Batteries and their Relationship to Item Difficulty," Unpublished paper, Mid-Continent Regional Educational Laboratory, Aurora, Colorado, 1987.

McPeek, John, *Critical Thinking and Education*, St. Martin's Press, New York, NY, 1981.

Modjeski, Richard B., and Michael, William B., "An Evaluation by a Panel of Psychologists of the Reliability and Validity of Two Tests of CT" *Educational and Psychological Measurement*, v43, n4, p1187-97, Winter 1983. [The tests reviewed were the "Watson-Glaser CT Appraisal" and the "Cornell CT Test."]

Morante E. A., and Ulesky, A. "Assessment of Reasoning Abilities," *Educational Leadership*, v42, p71-74, 1984.

Morgenstern, C. F., and Renner, J. W., "Measuring Thinking with Standardized Science Tests," *Journal of Research in Science Teaching*, v21, p639-648, 1984.

Nickerson, R. S., "Kinds of Thinking Taught in Current Programs," *Educational Leadership*, v42, p26-36, 1984.

Nickerson, R. S., *et al*, *The Teaching of Thinking*, Lawrence Erlbaum Associates, Hillsdale, NJ, 1985.

Norris, Stephen P. "Studies of Thinking Processes and the Construct Validation of CT Tests," in *Studies in CT Research Report No. 2*, Institute for Educational Research and.Development, Memorial University of Newfoundland, 48p, 1985.

_____, "Controlling for Background Beliefs When Developing Multiple-Choice CT Tests," Institute for Educational Research and Development, Memorial University of Newfoundland, Feb. 1988.

_____, "Informal Reasoning Assessment: Using Verbal Reports of Thinking to Improve Multiple-Choice Test Validity," in *Informal Reasoning and Education*, D. N. Perkins, J. Segal, and J. F. Voss (Eds.), Lawrence Erlbaum Publishing, Hillsdale, NJ, in press.

_____, "Verbal Reports of Thinking and Multiple-Choice CT Test Design," Institute for Educational Research and Development, Memorial University of Newfoundland, Apr. 1988.

Norris, Stephen P., and Ennis, Robert H. *Evaluating CT*, Midwest Publications, Pacific Grove, CA, 1987.

Norris, Stephen P., and Ryan, James, "Designing a Test of Inductive Reasoning," in *Proceedings of the International Conference on Argumentation 1986*, Van Eemeren, Frans H., and Grootendorst, Rob (Eds.), Foris Publications, Dordrecht, Holland, 1988 (?).

Paul, Richard W., "CT Research: A Response to Stephen Norris," *Educational Leadership*, v42, p46, 1985.

_____, "A New Theory of Knowing, Learning, and Literacy," Center for Critical Thinking and Moral Critique, Sonoma State University, 1988 (a).

_____, "35 Strategies for Infuins CT: Principles and Applications," Center for Critical Thinking and Moral Critique, Sonoma State University, CA, 1988 (b).

Pecorino, Philip, "CT Bibliography," *Newsletter on Teaching Philosophy*,, American Philosophical Association, p18, Summer 1987.

Quellmalz, Edys S. "Needed: Better Methods for Testing Higher-Order Thinking Skills," *Educational Leadership*, v43, p29-35, 1985.

Ruggiero, V. R., *Teaching Thinking Across the Curriculum*, Harper and Row, New York, 1988.

Scherer, Donald, and Facione, Peter A., "A Pre/Post Test for Introductory Logic Courses," *Metaphilosophy*, v8, n4, p342-347, Oct. 1977.

Sternberg, Robert J., *Beyond IQ: A Triarchic Theory of Human Intelligence*, Cambridge University Press, New York, 1985.

_____, (Ed.) *Teaching Thinking Skills: Theory and Practice* 1987.

Sternberg, Robert J., and Baron, Joan B., "CT: Its Nature, Measurement, and Improvement," National Institute of Education, Washington, DC, 37p, 1986.

Sternberg, Robert J., and Bhana, K., "Synthesis of Research on the Effectiveness of Intellectual Skills Programs: Snake-Oil Remedies or Miracle Cures," *Educational Leadership*, v44, p60-67, 1986.

Stewart, B. L., "Testing for CT: A Review of the Resources," *Rational Thinking Reports Number 2*, University of Illinois, Champaign-Urbana, IL, 1987.

Thompson, B., and Melancon, J. G., "Validity Measure of CT Skills," *Psychological Reports*, v60, p1223-1230. 1987.

Tucker, Carol B., "Examples of Question Types Used to Test Reasoning," Educational Testing Service, Princeton NJ, Apr. 1986.

_____, "Some Multiple-Choice Formats Used to Test Reasoning," Educational Testing Service, Princeton NJ, Apr. 1986.

_____, personal correspondence 3-9-88.

Woods, Dorothy A., *Test Construction*, Charles E. Merrill Books 1961.