ED 297 281 CS 009 233

AUTHOR Norris, Stephen P.

TITLE Informal Reasoning Assessment: Using Verbal Reports of Thinking to Improve Multiple-Choice Test Validity. Technical Report No. 430.

INSTITUTION Bolt, Beranek and Newman, Inc., Cambridge, Mass.; Illinois Univ., Urbana. Center for the Study of Reading.

SPONS AGENCY Social Sciences and Humanities Research Council of Canada, Ottawa (Ontario).

PUB DATE Jul 88

GRANT 410-83-0697

NOTE 24p.; For a related document, see CS 009 232.

PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.

DESCRIPTORS Cognitive Processes; *Critical Thinking; High Schools; High School Students; *Multiple Choice Tests; *Protocol Analysis; *Test Construction; Test Format; Test Items; *Test Theory; *Test Validity

ABSTRACT

A study examined whether the process of gathering verbal reports of subjects' thinking while taking multiple-choice critical thinking tests could be used to infer the reasoning process used and identify test items which do not require critical thinking skills. Four factors can render an inference of a subject's critical thinking skills untrustworthy: (1) the degree of informal reasoning sophistication of the subject; (2) the background empirical beliefs of the subject; (3) the assumptions which the subject brings to test items; and (4) the political and religious ideologies of the subject. Subjects, 343 senior high school students from four high schools, were divided into five groups for taking a multiple choice test. One group took the test in the normal manner while the other four groups gave verbal reports of their thinking for each question. Examiners asked subjects in these four groups leading questions to investigate whether the thinking processes of the subjects would be altered. Subjects were given a performance score equal to the number of correct answers and a "thinking score" which indicated the quality of thinking displayed in the verbal reports. Results indicated that there were no statistically significant differences in the performance scores of the five groups nor in the thinking scores of the four groups that used verbal reports. (One table of data and 30 references are attached.) (RS)

CENTER FOR THE STUDY OF READING

A READING RESEARCH AND EDUCATION CENTER REPORT

Technical Report No. 430

INFORMAL REASONING ASSESSMENT: USING VERBAL REPORTS OF THINKING TO IMPROVE MULTIPLE-CHOICE TEST VALID!TY

Stephen P. Norris
Memorial University of Newfoundland
and
University of Illinois at Urbana-Champaign

July 1988

University of Illinͻis at Urbana-Champaign
51 Gerty Drive
Champaign, Illinois 61820

The research upon which this publication was based was supported by a grant from the Social Sciences and Humanities Research Council of Canada, Grant No. 410-83-0697. It was completed while the author was a visiting scholar at the Center for the Study of Reading. The views expressed herein are those of the author and not the funding agency. I thank Robert Barcikowski, Robert Crocker, Susan Embretson, David Perkins, Linda Phillips, and Judith Segal for helpful comments. To appear as a chapter in D. N. Perkins, J. Segal, & J. F. Voss (Eds.), Informal reasoning and education. Hillsdale, NJ: Erlbaum.

Abstract

This paper examines some challenges to the validity of existing multiple-choice critical thinking tests and proposes how the validity of such tests might be put on a sounder footing. Several plausible hypotheses are proposed for explaining variance on critical thinking tests other than the hypothesis that the variance is due to differences in critical thinking. There is no evidence to support or rule out these alternative explanations. It is argued that asking samples of subjects to provide verbal reports of their thinking while working on such tests is a way to gather the needed evidence. The argument is supported by a study which showed that the thinking revealed in subjects' verbal reports while taking a test is likely an accurate representation of the thinking which they would have followed had they taken the test in its normal paper-and-pencil format.

INFORMAL REASONING ASSESSMENT: USING VERBAL REPORTS OF THINKING TO IMPROVE MULTIPLE-CHOICE TEST VALIDITY

A commonly understood characteristic of informal reasoning is that it can lead to multiple solutions for problems through multiple reasoning approaches. To accept such a state of affairs one does not have to be a complete anarchist, in the sense of being prepared to accept any solutions and any reasoning approaches. Restrictions can be made on the range of solutions and approaches that still leave room for considerable diversity.

Nevertheless, the possibility of diverse outcomes reached through diverse approaches creates problems for informal reasoning assessment. The problems are particularly acute when assessments involve the use of multiple-choice tests, since such tests reveal examinees' choices of answers but not the reasoning which led to those choices. If answers different from those keyed correct can be justified, then it is difficult to infer from examinees' answers alone the quality of their reasoning. If an examinee chooses the keyed answer, how proper is it to infer that some acceptable reasoning process was followed? Alternatively, if an examinee chooses an unkeyed response, how sound is the inference that an unacceptable reasoning process was followed?

Despite their shortcomings for informal reasoning assessment, multiple-choice tests are popular and likely to remain so. They are one major factor controlling instruction at the elementary and secondary school levels and indeed, one of the best means available for assessing some aspects of informal reasoning competence (Tomko & Ennis, 1980). This is not to say that multiple-choice tests can be used for all purposes. Essay tests, interviewing individual students, and direct classroom observation can serve purposes and yield information which multiple-choice tests cannot. For instance, all three seem better suited than multiple-choice tests to assessing informal reasoning *dispositions* (Norris & Ennis, in press). But using multiple-choice tests is probably the best way to develop student profiles on the many specific abilities which comprise informal reasoning, such as the ability to use the many criteria which are needed for judging the credibility of sources.

We are thus torn by two facts: (a) informal reasoning competence generally refers to the ability to use sound reasoning processes, rather than to the provision of adequate answers to tasks; and (b) multiple-choice tests, which provide no direct evidence on the reasoning processes used to accomplish tasks, are a popular and important approach for assessing informal reasoning competence. A question is whether existing multiple-choice tests of informal reasoning can adequately support inferences about the quality of reasoning processes and, if not, whether test construction practices can be improved so that future multiple-choice tests will be more valid.

This paper begins by challenging the validity of existing multiple-choice tests of informal reasoning. The point is made that the methodologies used to design these tests generally provide no direct evidence to counter the challenges. The second section proposes that eliciting verbal reports of thinking from examinees on trial test items is a way to obtain the direct evidence required. Research on the use of verbal reports in testing is sparse and provides little clear guidance on the usefulness of verbal reports for multiple-choice test validation. Some relevant research on verbal reporting from outside of testing is described, but there are still unresolved issues concerning the use of verbal reports of thinking for test validation. The third section reports a study designed to test the relevance of the evidence in verbal reports of thinking for validating multiple-choice tests of informal reasoning. The results suggest strongly that the evidence is relevant. Several implications for informal reasoning assessment are discussed in the final section.

Problems with Multiple-Choice Informal Reasoning Tests

When using multiple-choice tests of informal reasoning it is necessary to infer from the answers selected by examinees the reasoning processes they followed in reaching those answers. Several factors can render such inferences untrustworthy. The operation of four of these factors will be exemplified: the degree of informal reasoning sophistication of examinees; the background empirical beliefs of examinees; the assumptions which examinees bring to test items; and the political and religious ideologies held by examinees. The four factors overlap conceptually and empirically, but it is useful to distinguish them in this discussion in order to highlight different aspects of the overall problem of validating multiple-choice informal reasoning tests.

Degree of Sophistication of Examinees

Multiple-choice test items typically allow for only one correct answer. This restriction can create problems when testing reasoners with different degrees of sophistication in informal reasoning. By "different degrees of sophistication" I do not mean merely "different competence." A Grand Master is so much better than I at chess that comparisons of our competence are almost meaningless because we are in entirely different reference groups. It is this sort of difference that I am attempting to portray here, because the advertised audience for many multiple-choice informal reasoning tests is so broad as to make one wonder whether entirely different reference groups are being considered.

Let us examine an item from Section I of the Cornell Critical Thinking Test Level X (Ennis & Millman, 1985a), a popular multiple-choice test which assesses several aspects of informal reasoning competence. The test is aimed primarily at high school and undergraduate college students, but is recommended for use as low as fourth grade. Items are cast in the context of a story of a team of explorers that has just arrived on the newly discovered planet Nicoma. The explorers are searching for other explorers who arrived on Nicoma two years previously, but who have not been contacted since. Each item in Section I presents some information discovered by members of the second team and examinees are to decide whether the information is evidence for, evidence against, or neither evidence for nor against the hypothesis that all the members of the first team are dead. Here is the first item:

1. You go into the first hut. Everything is covered by a thick layer of dust.

The keyed answer is that the item presents evidence for the hypothesis that the members of the first group are all dead. However, judgments of the direction of evidence can vary legitimately with the informal reasoning sophistication of examinees. Suppose, reasoning in the following manner, an examinee concluded that the information in Item 1 was evidence neither for nor against the hypothesis that all the members of the first team are dead:

> I conclude that the information in Item 1 is evidence neither for nor against the hypothesis that all the members of the first team are dead. There are just too many ways to explain the information and we do not have sufficient information to choose among the possibilities. Maybe the first team stopped using this hut. Maybe they are using the hut for activity that raises a lot of dust. Maybe they have moved to another place on Nicoma. Maybe in fact they are all dead. Given that all of these possibilities can explain the information and given that there is insufficient information to choose among the possibilities, my theory of evidence says that the information is evidence neither for nor against *any* of the possibilities, including the hypothesis that all the members of the first team are dead.

There may be reason to disagree with the reasoning of this examinee. However, it is unlikely that the reasoning could be considered bad. In fact, the person's reasoning is quite sophisticated and it is this very sophistication which led to choosing an answer for Item 1 other than the one keyed correct. However, concurring with the key and marking the examinee's answer incorrect would not do justice to the level of the person's thinking. In a paper-and-pencil sitting where choice of answer is all that is recorded, this is exactly what would happen.

The same point can be illustrated using an item from the Cornell Critical Thinking Test Level Z (Ennis & Millman, 1985b), a test aimed at more sophisticated reasoners than Level X. The item is in Section II of the test and portrays two people debating whether or not the drinking water of Gallton ought to be chlorinated. Some thinking in the debate is faulty and, for each item, examinees are to choose from a list the best reason why the thinking is faulty. Here is the item:

> 11. DOBERT: I hear that you and some other crackpots are trying to get Gallton to chlorinate its water supply. You seem to think that this will do some good. There can be no doubt that either we should chlorinate or we shouldn't. Only a fool would be in favor of chlorinating the water, so we ought not to do it.
>
> ALGAN: You are correct at least in saying that we are trying to get the water chlorinated.
>
> Pick the one best reason why some of this thinking is faulty.
>
> A. Dobert is mistakenly assuming that there are only two alternatives.
>
> B. Dobert is using a word in two ways.
>
> C. Dobert is using emotional language that doesn't help to make his argument reasonable.

Alternative A appears to be true, since there are many alternatives, that range from not chlorinating at all to chlorinating using different concentrations of chlorine. Alternative B does not seem to be true. Alternative C, however, also appears true. There is thus a problem of deciding whether A or C offers the best reason for saying some of Dobert's thinking is faulty. The keyed answer is C on the grounds that, compared to the objection in C, it is insignificant to object that there are more than the two alternatives Dobert considers. However, a sophisticated informal reasoner might choose A on the grounds that it is Dobert's misunderstanding of chlorination which leads to his emotional outcry. The person might reason that if Dobert had an understanding that chlorination can occur in different degrees, then Dobert might have concluded that some level of chlorination is tolerable and not have become emotional. A sophisticated reasoner is more likely to see how people's beliefs, even about technical matters such as levels of chlorination, can affect their emotional responses. But this very sophistication can lead to being marked wrong on paper-and-pencil multiple-choice tests.

Problems can arise in other ways because of the different degrees of sophistication of examinees. Some items used to test for informal reasoning ask examinees to choose a level of endorsement for conclusions. However, examinees with different degrees of sophistication can justifiably choose different levels of endorsement, leading again to the possibility of examinees choosing unkeyed answers, even though they reasoned well. An example of such an item is found in the Watson-Glaser Critical Thinking Appraisal (Watson & Glaser, 1980), a test designed primarily for the junior high school level on up. In the item, examinees are to read a passage and assume that what it says is true. They then read a statement and judge, based on the informtion in the passage, whether it is True, Probably True, Probably False, or False, or that there is Insufficient Data to decide. The analysis which follows is based upon an analyses in Ennis and Norris (in press) and Norris and Ennis (in press).

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.

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.

The answer keyed correct is *Probably True*, which is said in the test manual to mean that it is more likely to be true than false that on some nights it was to Mr. Brown's advantage to keep his pool hall often after 1 a.m. However, a sophisticated informal reasoner might be able to imagine several alternative explanations of the facts. Mr. Brown might not have kept the pool hall open, but his son, whom Mr. Brown had recently put in charge of the business, kept it open. Mr. Brown was willing to take the blame and pay the fines for his son's offenses because he felt guilty for having neglected his son for many years. Maybe Mr. Brown had not kept the pool hall open, but had admitted he did in order that the fine could fall into the hands of corrupt municipal authorities as payment for giving him a license. Perhaps Mr. Brown had suffered a severe personal shock that resulted in his doing things which were not to his advantage. Perhaps Mr. Brown was protesting the discriminatory laws of his town which allowed some businesses to remain open later than 1 a.m., even though there were no principled reasons for doing this. He was protesting on principle, not because he thought the protest would be to his advantage. A sophisticated informal reasoner could conceive of possibilities such as these and, if a number of possibilities occur to a person when there is not enough information to adjudicate among them, then the person can justifiably choose *Insufficient Data*.

As another possibility, imagine a less sophisticated person who had learned that business people often break the law to their advantage, if the fines are small enough. Suppose the person also believes that a fine of $500 is sufficiently large that the only explanation of a business person's repeatedly acting so as to be levied such a fine is that the action is to the person's advantage. This examinee might justifiably choose *True*. Either way, examinees reasoning justifiably according to their level of sophistication would be marked wrong on paper-and-pencil sittings.

Background Empirical Beliefs of Examinees

Examinees bring different background beliefs to bear on multiple-choice informal reasoning tasks. The effect of such differences can be illustrated using a question from Section II of the Cornell Critical Thinking Test Level X. Recall that a team of explorers has landed on Nicoma to search for a team that has not been contacted in two years. The second team is exploring the area around their landing site and has found some water. In the item, the task is to choose which, if either, of two underlined statements is more believable.

27. A. The health officer says, "<u>This water is safe to drink</u>."
    B. Several others are soldiers. One of them says, "<u>This water is not safe</u>."
    C. A and B are equally believable.

The answer keyed correct is that the health officer's statement is more believable, because the health officer should be more expert than the soldier on the potability of water and because experts speaking in their own fields tend to be more believable than nonexperts. Suppose, however, an examinee believes that the training of soldiers and the equipment they carry equips them to make as dependable tests of water safety as health officers. Such an examinee would choose <u>C</u> as the answer, because the health officer and soldier are equally expert, but would be wrong according to the answer key. However, the examinee would have known that expertise in a field tends to make one more credible and would have used that criterion for choosing <u>C</u>. This is *precisely* the informal reasoning competence the item is designed to reward. But the person choosing <u>A</u> would be rewarded and the person choosing <u>C</u> penalized, even though the difference between them would have been their background empirical beliefs about the relative expertise of soldiers and health officers and not their informal reasoning competence.

Consider another example based on the Test on Appraising Observations (Norris & King, 1983). Items are set in the context of a traffic accident and various witnesses and people involved in the accident are reporting to police what they observed happening In Item 9, Ms. Vernon and Martine, two witnesses, are reporting on cars they had seen going through a stop sign. The task for examinees is to judge which of the underlined reports is more credible.

> 9. Ms. Vernon then says, "I also remember that a fancy blue sports car went through the stop sign."
> Martine says, "A car with twin headlights went right through the stop sign."

This item is designed to test the Principle of Observational Salience: Observations of more salient features of events tend to be more believable than observations of less salient features. Features of an event are salient to the extent that they are extraordinary, colorful, novel, unusual, and interesting and not salient to the extent that they are routine, commonplace, and insignificant. The answer keyed as correct, based on the empirical belief that being a fancy blue sports car is more salient than having twin headlights, is that there is more reason to believe Vernon's statement.

An examinee reasoning as follows would use the Principle of Observational Salience, but would not choose the answer keyed as correct.

> A fancy blue sports car is something which would stand out, but blue is not as noticeable a colour as red and there are a lot of fancy blue sports cars around nowadays. Twin headlights aren't as popular as they were in the past when just about every car had them, so they would stand out too. I believe neither would stand out more than the other, so the statements are equally believable.

This examinee knew the principle of informal reasoning being tested, but would have been marked wrong because of his or her empirical belief that having twin headlights is as salient a feature these days as being a fancy blue sports car.

Assumptions Made by Examinees

Different examinees make different assumptions while working on the same multiple-choice informal reasoning items. Moreover, there are different assumptions that can lead *justifiably* to different choices of answers. Consider the following items from the Interpretation subtest of the Watson-Glaser test. The task is to decide whether or not the numbered statements follow beyond reasonable doubt from the information given in the paragraph.

> Pat had poor posture, had very few friends, was ill at ease in company, and in general was very unhappy. Then a close friend recommended that Pat visit Dr. Baldwin, a reputed expert on helping people improve their personalities. Pat took this recommendation and, after three months of treatment by Dr. Baldwin, developed more friendships, was more at ease, and in general felt happier.
>
> 55. Without Dr. Baldwin's treatment, Pat would not have improved.
> 56. Improvements in Pat's life occurred after Dr. Baldwin's treatment started.
> 57. Without a friend's advice, Pat would not have heard of Dr. Baldwin.

The keyed answers are that the statement in Item 56 follows beyond reasonable doubt from the information in the paragraph and that the other two statements do not follow beyond reasonable doubt. In fact, the statement in Item 56 follows beyond *all* doubt, because the information includes the fact that the improvements occurred after three months of treatment by Dr. Baldwin. This indicates a serious problem with the items, because it seems the standards for being beyond *reasonable* doubt are taken by the test developers to be the same as those for being beyond all doubt.

However, imagine an examinee who understands "beyond reasonable doubt" in its everyday sense and ponders Item 55 as follows, making the assumptions indicated:

> The statement is ambiguous between "would not have improved ever" or "would not have improved during the three month period." It is obvious that there is insufficient information to say beyond reasonable doubt that Pat would never have improved without the help of Dr. Baldwin, so the statement must mean "would not have improved during the three month period." But is it beyond reasonable doubt that he would not have improved during this three month period had he not received Dr. Baldwin's treatment? Well, from the description, I assume that Pat had been suffering in this way for a long time. Problems such as this typically do not occur overnight, nor typically go away quickly by themselves without professional help. I therefore assume that Pat's problem was not one that would have gone away quickly on its own. Given these assumptions, the most plausible explanation of Pat's improved condition is that it was brought about by the treatment and therefore, while I cannot be certain, it seems beyond reasonable doubt that without Dr. Baldwin's treatment there would not have been such an improvement during the three months.

Such an examinee would be reasoning well, but would choose an answer other than the one keyed correct and be penalized for that in a paper-and-pencil sitting. The person made justifiable assumptions which were different from those of the test developers and these different assumptions, coupled with sound informal reasoning, led to the choice of an answer that would receive no credit in a paper-and-pencil sitting.

Ideologies of Examinees

Conceptions of informal reasoning competence do not incorporate or presuppose any political or religious ideology. Being subject to reason might be considered an ideology but, if so, it is not a political or religious one. However, political ideology can influence choices of answers on some informal reasoning items. Consider, for example, Items 65 and 67 from the Watson-Glaser test. Examinees are presented with the question, "Would a strong labor party promote the general welfare of the people of the United States?" Possible answers to the question and reasons defending those answers are provided:

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

Examinees are to assume the reasons are true and to decide whether they make strong or weak arguments for the answers given. They are told that strong arguments are those which are both important and directly related to the question.

Item 65 is keyed as giving a strong argument. However, for a laissez-faire examinee, the prospect of sustained large-scale unemployment might not be important compared to the interference required to suppress a labour party. So, although the argument in 65 might be directly related to the question, it is considered unimportant by the examinee and is, therefore, judged to be weak. On the other hand, a social activist examinee might also mark Item 65 as weak, but for different reasons. The person might, for example, believe that sustained large-scale unemployment would be a good thing because it would arouse the general public to revolt against the existing economic system. For this person, the reasons given in the item would not support the "No" answer to the question.

Item 67 is keyed as giving a weak argument. However, a political conservative might consider the argument both important and directly related to the question and, therefore, mark the item as strong. The conservative might believe that a strong labour party would encourage unions, which would lead to strikes in important industries, and believe that such strikes would be detrimental to the general welfare of the people of the United States. Given these beliefs the person could, while reasoning well, decide that the argument is strong.

Section Summary and Conclusion

Multiple-choice tests of informal reasoning provide only examinees' choices of answers to tasks, even though it is the reasoning which led to the choices and not the choices themselves that is of greatest interest. There is no direct evidence for the reasoning followed, so it must be inferred from the choices of answers. Several differences among examinees can make such inferences untrustworthy: different levels of informal reasoning sophistication; different background empirical beliefs; different assumptions made while taking tests; and different political and religious ideologies. This section has used items from exisiting multiple-choice informal reasoning tests to illustrate how each of these differences can lead to incorrect inferences about examinees' informal reasoning competence.

The items used to make these illustrations are not anomalies. They are indicative of a widespread problem in multiple-choice tests of informal reasoning. First, it is plausible that examinees with different levels of informal reasoning sophistication, different background empirical beliefs, and so on think differently about items. Second, there is no direct evidence (one way or another) of the extent to which such differences affect the trustworthiness of the inferences about examinees' reasoning.

Given the popularity and usefulness of multiple-choice informal reasoning tests, it is important to know whether anything can be done to increase their validity. A multiple-choice test of informal reasoning would be valid if, in general, good informal reasoning led to responses keyed correct and poor reasoning led to responses keyed incorrect. This condition for validity implies that evidence is needed on the relationship between the answers examinees choose and their reasoning. One plausible way to collect such evidence is to ask examinees to think aloud while working on trial items. Evidence gathered in this way has been espoused often but used rarely in validating multiple-choice informal reasoning tests. But a test founded on such evidence could resist strongly the criticisms posed in this section. Therefore, I shall turn to an exploration of the usefulness of verbal reports of thinking for improving multiple-choice informal reasoning tests.

Using Verbal Reports of Thinking to Validate Tests

Verbal reports of thinking contain information on the knowledge, strategies, and principles of reasoning which lead to examinees' choices of answers. They are not a means of observing directly reasoning processes, but verbal reports enable more trustworthy inferences about reasoning than just an examination of the answers chosen.

Using verbal reports of thinking goes hand in hand with the construction of theories of human mental abilities, by providing direct evidence for hypotheses about reasoning processes. The construct validation of ability tests has also been linked to theory construction (Cronbach, 1971), so it is natural to think that verbal reports of examinees' thinking are relevant to construct validation (Haney & Scott, 1987). If part of construct validation is the identification of the mental processes which underlie task performance, as argued by Embretson (1983) in her conception of *construct representation*, then the relevance of verbal reports of thinking to construct validation can be more readily seen. A multiple-choice informal reasoning test would have construct validity to the extent that good performance, defined in terms of the number of items answered correctly, could be explained by examinees' following sound thinking and poor performance could be explained by unsound thinking. Verbal reports of examinees' thinking while answering test questions can thus provide direct evidence for the construct validity of a test.

For verbal reports of thinking to be useful in the validation of an informal reasoning test, there must be a systematic procedure for collecting the reports, extracting information from them, and using that information for judging the quality of the test. More specifically, there needs to be a way to elicit verbal reports of examinees' reasoning that interferes with their reasoning as little as possible. There must be a means to use the information in the reports to judge examinees' reasoning independently of their answers to the test items, while being sensitive to different levels of sophistication of informal reasoning, different background beliefs, different assumptions, and different political and religious ideologies. Finally, there must be a way to compare answers chosen to the quality of reasoning and to determine the extent to which good and poor reasoning lead, respectively, to answers keyed correct and answers keyed incorrect.

There are several ways to elicit from examinees verbal reports of their thinking on test items. They might be asked simply to say everything that comes to their minds as they work on a task. Alternatively, they might be asked to justify their answers. Examinees might be probed with questions about the specifics of their reasoning, by being asked whether such-and-such had anything to do with their thinking and, if so, what role it played. Finally, some combination of these approaches might be used.

Whatever the specifics, it is not clear whether different elicitation approaches yield more or less the same information on examinees' reasoning, or whether any approach yields trustworthy information on thinking. But for a test validation methodology to rely on verbal reports of thinking, these issues must be clarified. It is not sufficient to argue, as I have done so far, that in principle verbal reports of thinking ought to be relevant to the validation of multiple-choice tests of informal reasoning.

In reality, verbal reports of thinking are relevant to the validation of multiple-choice informal reasoning tests, only if the information on examinees' thinking which the reports contain is an accurate reflection of the thinking which would have taken place had the examinees taken the test in normal paper-and-pencil format. Verbal reports of thinking require that subjects provide introspective reports on the progress of their thinking or the reasons for their performance, often in the presence of an investigator. It is not known how such requirements influence thinking and the small number of testing studies which have used verbal reports of thinking (Bloom & Broder, 1950; Connolly & Wantman, 1964; Kropp, 1956; McGuire, 1963; Schuman, 1966) have ignored the question. There is some relevant research from non-testing contexts, such as information processing research on the use of verbal reports as data and memory research on eyewitness testimony. I will thus briefly review the research in each of these areas.

Verbal Reports as Data

Research on the trustworthiness of verbal reports of mental processes points to conflicting conclusions. On the one hand, Nisbett and Wilson (1977) conclude that people have little or no introspective access to the things which stimulate their cognitive processes. On the other hand, Ericsson and Simon (1980, 1984) and Smith and Miller (1978) claim that people do have dependable access to their mental processes *in certain situations*.

To support their conclusion Nisbett and Wilson reviewed evidence from the cognitive dissonance, self-perception attribution, learning without awareness, and problem-solving literatures. Based upon this evidence, they conclude three things: (a) people often cannot accurately report the effects of certain stimuli on their responses to problems requiring higher order thinking; (b) when people do report on such stimuli they often do not search their memories to discover what the stimuli were, but rather appeal to plausible hypothetical mechanisms which they accept a *priori*; and (c) when people are correct about the stimuli affecting their responses they have coincidentally appealed to a hypothesis which happens to be correct. Nisbett and Wilson argue that such coincidences occur when the actual causal stimulus is available to memory because, *a priori*, it is the most plausible cause of the response.

Smith and Miller take issue with these conclusions, because the experimental situations upon which the conclusions are based do not support the generalizations made in them. In particular, experiments are situations in which the influential stimulus is "systematically and effectively [hidden] from [subjects] by [the] experimental designs" (1978, p. 356). The influential stimulus can only be ascertained by comparing the treatment and control groups and, of course, subjects in an experiment cannot do this. Therefore, Smith and Miller argue that Nisbett's and Wilson's conclusions apply only to experimentally controlled situations in which subjects' unawareness of what is influencing their thinking is a natural consequence of the experimental setup. They claim that these experimental findings are not generalizable to people's attempts to report on their mental processes outside of experimental settings. Reports of thinking on test items might thus fall outside the scope of Nisbett's and Wilson's conclusions, because testing does not usually attempt to hide influential stimuli from examinees.

Ericsson and Simon (1980, 1984) discuss the trustworthiness of verbal reports of thinking from an information processing perspective. They conclude that instructions to verbalize slow down, but do not change the course of, cognitive processing when subjects are verbalizing information that would normally be available to them in short-term memory. Specific and directive probes alter cognitive processing, however, as do requests to supply motives and reasons. This conclusion is particularly relevant for test validation contexts where reasons for answers might be sought. The conclusion suggests that some verbal reports of thinking on test items might not be applicable to testing contexts in which verbal reporting is not done.

Ericsson and Simon make specific hypotheses about how different types of requests to think aloud can affect the trustworthiness of verbal reports of thinking. In particular, they hypothesize that the less leading the probe employed the more accurate the information obtained, and that more information with an overall lower trustworthiness can be obtained with more leading probes. These hypotheses need to be tested.

It is not legitimate to assume that the research on verbal reports as data answers all the questions relevant to the use of verbal reports of thinking in testing situations. Testing contexts are sufficiently different from experimental and information processing research contexts that it is reasonable to expect that memory retrieval and information processing demands might also differ. In particular, test-takers make specific assumptions about how they should try to perform and how the results reflect upon them that are probably different from those made when involved in a psychological study.

Eyewitness Testimony Research

Eyewitness testimony is often contained in verbal reports given in response to questions. Verbal reports of thinking on tests are similar sorts of things. In one situation, people try to remember what they have observed; in the other, they try to remember what they have thought. The remembering processes are likely related, though not identical. Thus, research on the factors which affect the accuracy of eyewitness testimony is pertinent to the question of the accuracy of verbal reports of thinking on tests. The degree of pertinence is tempered by dissimilarities between the two contexts: in one, the memory is of an external event, whereas in the other it is of an internal event; in one, the memory is of events in the more distant past, whereas in the other the memory is of events in the very recent past.

The eyewitness testimony research most relevant to the present study explores the effect of different of questioning on the accuracy of observation reports. Three categories of questions have been (Loftus, 1979, p. 90): (a) those eliciting *free* reports (e.g., "Tell us all that you saw"); (b) those *controlled* reports (e.g., "Give us a description of what our assailant was wearing"); and (c) those eliciting *alternate-choice* reports (e.g., "Did your attacker have dark or light hair?"). Two general conclusions can be drawn on the basis of many independent studies of these types of questioning techniques (Clifford & Scott, 1978; Dale, Loftus, & Rathbun, 1978; Harris, 1973; Hilgard & Loftus, 1979; Lipton, 1977; Loftus & Palmer, 1974; Marquis, Marshall, & Oskamp, 1972). The first conclusion is that free reports tend to be more accurate than any other type of report; controlled reports rank next in accuracy; and alternate-choice reports have the lowest degree of accuracy. The second conclusion is that the amount of information obtained increases in the opposite direction: free reports contain the least amount of information; controlled reports contain somewhat more information; and alternate-choice reports contain the most information of all. So then, free reports give a relatively lesser amount of relatively more accurate information, and alternate-choice reports a relatively greater amount of relatively less accurate information. The results are consistent with the hypotheses offered by Ericsson and Simon (1980, 1984).

As with the research on verbal reports as data, it is not legitimate to assume that the results of eyewitness testimony research can be applied directly to testing. Eliciting reports of thinking on tests is different from eliciting recollections of observed events and there is no research which explores how these differences affect the accuracy of both types of report.

An Unresolved Problem

Let us retrace. The evaluation of informal reasoning competence makes demands which traditional multiple-choice tests are not equipped to meet. Problems requiring informal reasoning for their solution often admit of more than one solution, but multiple-choice tests usually have only one correct answer. Evaluators of informal reasoning are usually more interested in the process of examinees' reasoning than the product, but multiple-choice tests typically give no direct evidence on reasoning processes.

Despite these problems, multiple-choice tests are likely to continue to be used and to have considerable influence. Therefore, it would be worthwhile to have a way to validate the tests which can provide some direct evidence on the reasoning processes they elicit. One natural way to gain direct evidence on reasoning is to ask people to think aloud. Applied to the validation of multiple-choice informal reasoning tests, tests could be examined by asking samples of examinees to work on them and to report verbally on their thinking. Judgments could be made of whether or not good and poor informal reasoning led, respectively, to keyed and unkeyed answers. Specifically, the evidence could indicate whether differences in performance across an intended audience for the test was significantly affected by such factors as differences in reasoning sophistication, background empirical beliefs, assumptions made, and religious or political ideologies.

The idea is sound in the abstract. But there is still much to learn about how thinking aloud affects thinking itself. More particularly, there is virtually no research on the use of verbal reports of thinking in testing contexts, and the verbal reports as data and eyewitness testimony literatures can be taken only as suggestive of what to expect in testing. The use of verbal reports of thinking to validate tests would be justified, only if their elicitation does not alter significantly the course of examinees' thinking from what it would have been had they worked on the tests in paper-and-pencil format. If a significant alteration occurs, then information on the validity of tests derived from the verbal reports of thinking would not provide evidence on the validity of the tests for the paper-and-pencil sittings for which most are intended. It is therefore worth exploring whether verbal reports of thinking on multiple-choice informal reasoning tests provide relevant evidence on the validity of those tests.

Relevance of the Evidence in Verbal Reports of Thinking

The issue of the relevance of verbal reports of thinking to validating multiple-choice tests of informal reasoning was studied by trying to answer two research questions:

1. Do different ways of requesting verbal reports from examinees yield different information on their thinking?
2. Does the act of verbally reporting thinking alter examinees' test performance?

The first question pertains to the role of the interview procedure. As stated earlier, slight changes in the wording of interrogations of eyewitnesses can cause different accounts of events to be given. Is the same true when asking examinees to verbally report their thinking? The second question addresses the issue of how verbally reporting one's thinking alters the course of that thinking. If significant alterations occur, they should be revealed in different test performances between examinees who verbally report their thinking and those who do not.

Description of the Study

To help answer these questions, 343 senior high school students from four high schools participated in an experiment. Verbal reports of their thinking were elicited as they worked through Part A of the Test on Appraising Observations (Norris & King, 1983). As described previously, it is a multiple-choice test focussed on one aspect of informal reasoning competence, the ability to judge the credibility of reports of observations. In Part A, items are cast in the context of a traffic accident. In each item two people, either witnesses or individuals involved in the accident, provide accounts of what happened. Examinees are to judge which, if either, of the accounts is more credible. Judgments should be based on characteristics of either the observers, the observation conditions, or the statement of observation itself.

A completely randomized factorial design was used. Students were randomly assigned to one of five groups:

1. **No Probe Group:** Students were not interviewed and worked alone on the test in a paper-and-pencil format.

2. **Think Aloud Group:** Students were instructed to report all they were thinking as they worked through the items.

3. **Immediate Recall Group:** Students were asked to choose their answers to each question and to justify their choices immediately after each was made.

4. **Criteria Probe Group:** While working on each question, students' attention was drawn to a particular piece of information in it. They were asked whether that information made any difference to the answers they chose and, if so, to explain the difference.

5. **Principle Probe Group:** Students were treated as in the criteria probe group, except they were asked an additional question aimed at determining whether their choices were based upon particular general principles.

The no probe group simulated conditions under which the test would normally be given. Students worked alone at a desk and marked their answers on an answer sheet. In the think aloud group there was considerable leeway for students to think and report as they saw fit, because only the general instruction to think aloud was given. In subsequent groups, students responses were constrained by leading requests for particular sorts of information. The types of probes vary analogously in degree of leadingness to tbose studied in eyewitness testimony research. If the results of that research generalize to testing situations, then students' verbal reports of thinking should vary depending upon their probing group.

Let us consider how the system would proceed for each of the groups working on a given item. Here is Item 3 as an example:

> A policewoman has been asking Mr. Wang and Ms. Vernon questions. She asks Mr. Wang, who was one of the people involved in the accident, whether he had used his signal.
> Mr. Wang answers, "Yes, I did use my signal."
> Ms. Vernon had been driving a car which was not involved in the accident. She tells the officer, "Mr. Wang did not use his signal. But this didn't cause the accident.

Students were to choose which, if either, of the underlined statements is more credible. In addition, the following instructions were given to students in each interviewed group:

Interviewed Group	**Instructions to Examinees**
Think Aloud	Try to tell me all that comes to your mind as you think about this question.
Immediate Recall	Which answer do you choose? . . . Can you tell me why you chose that answer?
Criteria Probe	Which answer do you choose? . . . Did the fact that Mr. Wang was involved in the accident affect your choice?
Principle Probe	Which answer do you choose? . . . Did the fact that Mr. Wang was involved in the accident affect your choice? . . . (If "No") Go on to the next item. (If "Yes") What difference did it make to your thinking that he was involved?

Students' verbal reports were tape recorded and transcribed verbatim. All students were assigned *Performance Scores* equal to the number of items answered correctly according to the key provided with the test (Norris & King, 1985). Students who had given verbal reports were also assigned *Thinking Scores*. These scores indicated the quality of thinking displayed in the verbal reports on a scale of 0-3 for each item. They were assigned independently of answers chosen.

Quality of thinking was judged by comparing students' verbal reports to ideal models of thinking developed for each item. The models were based upon a set of principles for assessing the credibility of observations, knowledge of which the test was designed to measure. Staying with Item 3, the ideal model was based as follows on the principle that people in a conflict of interest tend to be less credible than those not in a conflict of interest:

> Mr. Wang was involved in the accident, but Ms. Vernon was not involved. Mr. Wang is less credible because his involvement would give him reason to say he used his signal even if he did not. Wang is in a conflict of interest. People in a conflict of interest, that is, people who have something to gain by events being cast as they described them, tend to be less credible than those who are not in such a situation.

According to the model, an examinee first needs to identify in the text the relevant information about Wang's and Vernon's involvement. The text is simple enough that reading ability should not impede this identification for most high school students. Second, an examinee must remember from experience that not using a turn signal can cause an accident and that being held responsible for an accident can be troublesome. High school students should have ready access to such common knowledge. Finally, an examinee has to recognize that being in a conflict of interest is an accuracy-reducing factor and apply this principle to make a credibility judgement.

So for Item 3, thinking scores were assigned according to the following scale:

1 point:	The examinee points out that Mr. Wang was involved in the accident.
2 points:	The examinee points out that Mr. Wang was involved in the accident and compares Mr. Wang's involvement to Ms. Vernon's being a bystander.
3 points:	The examinee points out that Mr. Wang was involved in the accident, compares this with Ms. Vernon's non-involvement, and shows that this is an instance of a more general phenomenon in which people stand to profit or lose depending upon what they say.
0 points:	The examinee does none of the above or does not respond.

Generalizing to all items, students were assigned one point towards their thinking scores for each of the following:

1. citing the relevant facts in the text which can be used to compare the underlined statements for their credibility;

2. using these facts together with any relevant background knowledge to make a comparative evaluation of the credibility of the statements;

3. showing how the evaluation is based on a generalized accuracy-reducing factor.

To illustrate the procedure more clearly, let us examine a transcript of one student's verbal report of thinking on Item 3. The student said:

> The second one 'cause, ah, 'cause he'd say that he used the signal so he wouldn't have nothing to do with the accident. Probably afraid he'd have . . . he'd be questioned by the police or something.

Note that verbal reports of thinking tend to have many colloquialisms, repetitions, and false starts. This is how we speak and these things must be overlooked when rating examinees' thinking. This student would be assigned a thinking score of 2. There is judgement involved in this decision, because there is no exact one-to-one correspondence between what the student said and the rating scale above. But the student clearly recognized the accuracy-reducing role of Wang's being involved in the accident. The student did not cite explicitly the facts that Wang was involved and Vernon was not, but these were clearly implicit in the student's thinking. The student would not be given a 3, because no general principle was cited.

Results

The verbal reports of thinking, the thinking scores, and the performance scores were analzyed quantitatively and qualitatively in an attempt to answer the two questions raised at the beginning of the previous section (for more details see Norris, 1985):

1. Do different ways of requesting verbal reports from examinees yield different information on their thinking?

2. Does the act of verbally reporting thinking alter examinees' test performance?

The results of the quantitative analysis of thinking scores showed no statistically significant differences across the four groups that were interviewed. So in answer to the first question, whatever other effects the different types of probes might have had, they did not affect the quality of students' thinking as measured by the thinking score scale.

To further examine the thinking of students in the different interview groups, a qualitative analysis was conducted of a random sample of 40 (stratified by interview group) of the total sample of 271 interviews. A coding scheme was devised for indicating a variety of verbal moves in the examinees' verbal reports. The moves are as follows:

Citing Factual Details - either recalling a factual detail given in an item prior to the one currently being done, recalling such a prior detail incorrectly, or stating a detail in the current item;

Asking Rhetorical Questions - posing questions which appear to be directed to the examinee himself or herself rather than to the interviewer;

Making Evaluations - either evaluating judgments or conclusions which had been previously explicitly stated, or evaluating ones which had not been verbalized;

Constructing Supporting Assumptions - either making detailed factual assumptions specific to the current item, or making more generalized assumptions of broad principles of appraisal or causal laws covering more than the situation in the current item;

Using Attention Control Devices - either making comments about the stage of progress reached in reasoning through the problem (Let's see. Where was I?), or commenting on the direction reasoning should proceed (Wait now!);

Interacting with the Experimenter - directing comments or questions to the experimenter;

Pausing - either making verbal inflections (Ahhh! Mmmm!) or being silent.

The 40 verbal reports of thinking were coded according to the seven categories and frequencies of verbal moves calculated. These frequencies are given in Table 1. No systematic analysis was done on these data, but they were examined for general trends with a view to more systematic exploration in the future. Note that there are clear differences in frequencies of occurrence among verbal move categories. However, there are no glaring differences in trends across the interview groups, supporting the conclusion of the quantitative analysis that there was no difference in quality of thinking across the four interviewed groups.

[Insert Table 1 about here.]

Both the quantitative and qualitative results suggest strongly that it was subjects' thinking and not how that thinking was elicited that controlled what was reported. If this conclusion can be substantiated in other studies and for other tests, then it would seem that the accuracy of verbal reports of thinking on multiple-choice informal reasoning tests is not as sensitive to the type of probing as research in other contexts would indicate. That is, testing may be a context whose demands are sufficiently unique that the use of verbal reports of thinking on tests needs study in its own right.

The second question asked whether the act of verbally reporting thinking alters examinees' test performance. Analysis showed that there are no statistically significant differences in performance scores between any of the interviewed groups and the group who took the test in the paper-and-pencil format. This result suggests that probing did not alter thinking, because if the course of examinees' thinking had been altered by giving verbal reports on their thinking, then this alteration should have been revealed in altered performance. It is hard to imagine how their thinking could have differed in a systematic fashion while their performances stayed precisely the same.

Discussion and Conclusions

Whenever no differences between treatments is the result of an experiment, the power of the experiment to detect differences which actually exist becomes an important concern. Was this experiment sufficiently powerful to detect any differences which existed among the groups? There are a number of reasons to believe that differences would have been detected had they been present in the population. First, the treatments were considerably different from one another. It is quite different for high school students to work alone on a test than to work in the presence of a stranger who is probing their thinking. Thus, if eliciting verbal reports of thinking tend to alter the course of thinking, then alterations should have been revealed in differences in performance between the interviewed and uninterviewed groups.

Second, the interview treatments themselves were considerably different. The leading probes were quite leading, because they made explicit suggestions to students about what could have affected their choices of answers. It would have been an easy matter for students to conform to these suggestions by altering their answer choices and their way of thinking about items. Instead, students denied regularly that a suggested factor had anything to do with their thinking and proceeded to explain how their choices were made. That is, students tended to maintain whatever interpretation made sense to them.

Third, effects were sought from a number of different directions, but were found in none of them. The quantitative analysis showed no differences in thinking scores across the four interviewed groups and no differences in performance scores across all five groups. The qualitative analysis showed that the same patterns of verbal moves were used by students in each of the interviewed groups. It is plausible to think that if the verbal reporting altered students' thinking it would have been detected by at least one of these methods.

Fourth, eyewitness testimony research uncovers consistent effects using similar sorts of treatments. This does not mean that differences should have been found in this study, but it does mean that *if* differences existed they should have been detected.

Finally, an analysis of the statistical power of the experiment showed less than a 5% chance that real differences existed among the groups but were not detected.

This research points to a useful technique for validating multiple-choice tests of informal reasoning. Eliciting verbal reports of examinees' thinking is a plausible way to gather data on the quality of tests. This study bolsters confidence in the technique by showing that there is no need to be overly cautious about the leadingness of questions used to elicit reports of thinking. Examinees' thinking is not altered by requests to report on their thinking, so the information in the reports is relevant evidence for the validity of tests. Such evidence can show whether sophistication, background empirical beliefs, ideologies of reasoners, assumptions reasoners make, and other factors affect performance on multiple-choice informal reasoning tests.

Collecting verbal reports of thinking on existing multiple-choice informal reasoning tests should therefore provide important evidence on the validity of those tests. Given the level of suspicion cast on them by the sorts of criticisms discussed earlier. such evidence is needed. It is important to know, one way or the other, whether or not existing multiple-choice informal reasoning tests are valid.

The results of such validation studies might be mixed. For instance, whereas many multiple-choice informal reasoning tests are advertised for wide ranges of audiences, verbal reports of thinking from subjects across the entire range may indicate that the advertised applicability of a given test should be narrower. As a consequence, the advertised range of applicability might be altered or, using the information in the verbal reports of thinking, versions of a given test suitable for more narrowly defined audiences might be designed. These versions may differ considerably from each other, or may only differ in keyed responses. It might be possible, for instance, to tailor answer keys to different audiences to take account of such factors as sophistication, empirical beliefs, ideologies, and so on. As far as I know, this approach has never been tried with multiple-choice informal reasoning tests, but the information in verbal reports of examinees' thinking could provide a basis for such trials.

The idea of using verbal reports of thinking to tailor answer keys to different audiences suggests a developmental (in addition to validation) role for verbal reports. There is no reason to wait until tests have been developed before using verbal reports of thinking to check their validity. Verbal reports of thinking on trial items of a test under development can provide evidence for retaining, modifying, or discarding items. With a systematic procedure for quantifying and using this evidence to judge individual items and the test as a whole (Norris, 1988), validity can be "built into a test" from the item level on up. Verbal reports of thinking thus open the prospect of developing valid multiple-choice tests to do the sorts of informal reasoning assessment for which they are most suited.

However, not all informal reasoning assessment can be served by multiple-choice testing. The Test on Appraising Observations, used as an example in this chapter, assesses the ability to apply criteria one at a time to make appraisals of credibility. But in a real-world context of appraising the credibility of a witness, several of the criteria would likely apply at once. Some of the criteria might push the appraisal in one direction, others in another direction. The criteria would have to be weighed and balanced and there are no strict rules for doing this. Judgement based upon experience would have to be used. Multiple-choice tests are not useful for assessing how well people use their judgement to *orchestrate* a number of informal reasoning skills to work on ill-defined, real-world problems. Other assessment methods must be developed.

Informal reasoning *dispositions* also pass through the mesh of multiple-choice informal reasoning tests, but reasoning dispositions are as important to assess as reasoning abilities. The assessment of dispositions is logically a two-stage process, because failure to perform well (e.g., to give alternative hypotheses when appropriate) could be explained either by lack of knowledge that giving alternatives is appropriate, lack of ability to generate alternatives, or lack of disposition (given the knowledge) to provide alternatives. The possibilities of lack of knowledge and ability must be ruled out before lack of disposition can be accepted as the explanation. Assessment of dispositions is doubly complex and there are no adequate techniques for assessing dispositions to be open-minded, to seek reasons, to seek alternatives, to seek critical feedback, and so on. Furthermore, it is not clear at this time how these assessments might be done best. Essay testing, interviewing individuals, and direct classroom observation are approaches with promise (Norris & Ennis, in press), but considerable research is needed.

While the problems of informal reasoning assessment are large, they are surmountable. Many problems stem from the fact that educators have only recently taken seriously instruction in reasoning. Assessment practices which are adequate for goals of instruction that focus primarily on learning factual knowledge are not adequate for assessment of informal reasoning. Therefore, because of the new-found goal of teaching reasoning, many assessment practices will have to change, some will have to go, and new practices will have to take their place. This chapter showed why changes are needed and how changes can be made to multiple-choice testing to make it more suitably meet the goals of informal reasoning assessment.

References

Bloom, B. S., & Broder, J. L. (1950). *Problem-solving processes of college students.* Chicago: The University of Chicago Press.

Clifford, B. R., & Scott, J. (1978). Individual and situational factors in eyewitness testimony. *Journal of Applied Psychology, 63,* 352-359.

Connolly, J. A., & Wantman, M. J. (1964). An exploration of oral reasoning processes in responding to objective test items. *Journal of Educational Measurement, 1,* 59-64.

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

Dale, P. S., Loftus, E. F., & Rathbun, L. (1978). The influence of the form of the question on the eyewitness testimony of preschool children. *Journal of Psycholinguistic Research, 7,* 269-277.

Embretson (Whitely), S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93,* 179-197.

Ennis, R. H., & Millman, J. (1985a). *Cornell critical thinking test, level X.* Pacific Grove, CA: Midwest Publications.

Ennis, R. H., & Millman, J. (1985b). *Cornell critical thinking test, level Z.* Pacific Grove, CA: Midwest Publications.

Ennis, R. H., & Norris, S. P. (in press). Critical thinking testing and other critical thinking evaluation: Status, issues, needs. In J. Algina (Ed.), *Issues in evaluation.* New York: Ablex.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review, 87,* 215-251.

Ericsson, K. A., & Simon, H. A. (1984). *Protocol analysis: Verbal reports as data.* Cambridge, MA: MIT Press.

Haney, W., & Scott, L. (1987). Talking with children about tests: An exploratory study of test item ambiguity. In R. O. Freedle and R. P. Duran (Eds.), *Cognitive and linguistic analyses of test performance* (pp. 298-368). Norwood, NJ: Ablex.

Harris, R. J. (1973). Answering questions containing marked and unmarked adjectives and adverbs. *Journal of Experimental Psychology, 97,* 399-401.

Hilgard, E. R., & Loftus, E. F. (1979). Effective interrogation of the eyewitness. *The International Journal of Clinical and Experimental Hypnosis, 27,* 342-357.

Kropp, R. P. (1956). The relationship between process and correct item responses. *Journal of Educational Research, 49,* 385-388.

Lipton, J. P. (1977). On the psychology of eyewitness testimony. *Journal of Applied Psychology, 62,* 90-95.

Loftus, E. F. (1979). *Eyewitness testimony. Cambridge, MA: Harvard University Press.*

Loftus, E. F., & Palmer, J. C. (1974). Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of Verbal Learning and Verbal Behavior, 13,* 585-539.

Marquis, K. H., Marshall, J., & Oskamp, S. (1972). Testimony validity as a function of question form, atmosphere, and item difficulty. *Journal of Applied Social Psychology, 2,* 167-186.

McGuire, C. (1963). Research in the process approach to the construction and analysis of medical examinations. *National Council on Measurement in Education Yearbook, 20,* 7-16.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84,* 231-259.

Norris, S. P. (1985). *Studies of thinking processes and the construct validation of critical thinking tests.* St. John's, Newfoundland: Institute for Educational Research and Development, Memorial University of Newfoundland. (ERIC Document Reproduction Service No. ED 264 259)

Norris, S. P. (in preparation). *Using verbal reports of thinking to develop and validate multiple-choice critical thinking tests.* Manuscript submitted for publication.

Norris, S. P., & Ennis, R. H. (in press). *Evaluating critical thinking.* Pacific Grove, CA: Midwest Publications.

Norris, S. P., & King, R. (1983). *Test on appraising observations.* St. John's, Newfoundland: Institute for Educational Research and Development, Memorial University of Newfoundland.

Norris, S. P., & King, R. (1985). *Test on appraising observations manual.* St. John's, Newfoundland: Institute for Educational Research and Development, Memorial University of Newfoundland.

Schuman, H. (1966). The random probe: A technique for evaluating the validity of closed questions. *American Sociological Review, 31,* 218-222.

Smith, E. R., & Miller, F. D. (1978). Limits on perception of cognitive processes: A reply to Nisbett and Wilson. *Psychological Review, 85,* 355-362.

Tomko, T. N., & Ennis, R. H. (1980). Evaluation of informal logic competence. In J. A. Blair & R. Johnson (Eds.), *Informal logic: The first international symposium.* Inverness, CA: Edgepress.

Watson, G., & Glaser, E. M. (1980). *Watson-Glaser critical thinking appraisal.* Cleveland, OH: The Psychological Corporation.

Table 1

Frequency of Verbal Moves by Interview Group

	Interview Group			
Verbal Moves	Think Aloud	Immed. Recall	Crit. Probe	Princ. Probe
Citing Factual Details	104	139	99	139
Asking Rhetorical Questions	16	9	2	5
Making Evaluations	45	24	39	43
Constructing Assumptions	178	228	214	227
Attention Control	26	25	15	19
Interact with Experimenter	19	9	12	13
Pausing	499	387	424	380