ED 264 259 TM 850 741

AUTHOR	Norris, Stephen P.
TITLE	Studies of Thinking Processes and the Construct Validation of Critical Thinking Tests. Studies in Critical Thinking Research Report No. 2.
INSTITUTION	Memorial Univ., St. John's (Newfoundland). Inst. for Educational Research and Development.
SPONS AGENCY	Social Sciences and Humanities Research Council of Canada, Ottawa (Ontario).
PUB DATE	85
GRANT	410-83-0697
NOTE	48p.; An earlier draft was presented at the Annual Meeting of the American Educational Research Association (69th, Chicago, IL, March 31-April 4, 1985).
PUB TYPE	Reports - Research/Technical (143) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Analysis of Variance; *Cognitive Processes; *Cognitive Tests; *Critical Thinking; Data Analysis; Foreign Countries; High Schools; Interviews; Item Analysis; Research Methodology; *Response Style (Tests); Speech Communication; Test Interpretation; *Test Validity; Test Wiseness
IDENTIFIERS	Canada; *Test on Appraising Observations (Morris and King)

ABSTRACT

The usefulness of studying thought processes in the construct validation of ability tests was examined using a sample consisting of 343 Canadian senior high school students. Four levels of probing were used by the interviewer to examine the students' thinking processes while taking the Test on Appraising Observations: (1) think aloud; (2) immediate recall--examinees were asked why they chose that answer; (3) criteria probe--examinees were asked whether a feature of a test item determined the answer; and, (4) principle probe--examinees were also asked whether answer choice was based on particular general principles. Two scores were derived: performance scores or number right, and thinking scores indicating the quality of thinking displayed. Analyses were concerned with three questions: (1) whether the verbal reports accurately portray the thinking that takes place; (2) whether thinking concurrent with reporting differs from thinking in testing situations in which reports are not elicited; and (3) whether thinking subsequent to reporting is different from thinking subsequent to non-reporting testing situations. Analyses of variance were conducted to determine the effects and interactions of interview group, interviewer, sex, and grade level. Seven types of verbal activity were noted during the interviews. (GDC)

STUDIES OF THINKING PROCESSES AND THE CONSTRUCT VALIDATION OF CRITICAL THINKING TESTS

by

Stephen P. Norris
Institute for Educational Research and Development

INSTITUTE FOR EDUCATIONAL RESEARCH AND DEVELOPMENT
MEMORIAL UNIVERSITY OF NEWFOUNDLAND

STUDIES OF THINKING PROCESSES

AND THE

CONSTRUCT VALIDATION OF CRITICAL THINKING TESTS

Stephen P. Norris

Studies in Critical Thinking
Research Report No. 2

Institute for Educational Research and Development
Memorial University of Newfoundland
St. John's, Newfoundland
Canada
A1B 3X8

This research was supported by a grant from the Social Sciences and Humanities Research Council of Canada, Grant No. 410-83-0697. The views expressed herein are neither those of the Institute for Educational Research and Development nor those of the funding agency. An earlier draft of this manuscript was presented at the annual meeting of the American Educational Research Association, Chicago, April 1985. I thank Robert Barcikowski, Robert Crocker, Susan Embretson, and Linda Phillips for helpful comments.

This report describes an experiment designed to explore the usefulness of studies of thinking processes in the construct validation of ability tests. A study of thinking processes is one in which an attempt is made to gain information on the mental processes which people use to perform tasks, that is, to describe the strategies, and kinds of information which lead to performance. A study of thinking processes typically does not lead to the direct observation of mental processes (though the possibility of direct observation cannot be ruled out, especially in the future), but allows more trustworthy inferences to be made about their nature than can be obtained through the examination of performance at the strictly task level. A study of thinking processes represents a concerted attempt to "look beneath the surface" of directly observable task behaviour to discover its underlying causes. This requires both the invention and justification of appropriate probing techniques and the imaginative hypothesizing of mechanisms and processes which can account for what is found using these techniques.

Under the description above, studies of thinking processes go hand in hand with the construction of theories of human mental abilities, and are explicitly designed to facilitate this activity by increasing the reliability of inferences from data to theory. The process of construct validation of ability tests has also been linked to theory construction, so it is natural to think that studies of thinking processes are relevant to construct validation. If construct validation is conceived (at least in part) as the identification of the mental processes which underlie task performance, as has been done by Susan Embretson(Whitely) (1983) in her conception of construct representation, then the relevance of studies of thinking processes to construct validation can be more readily seen. Evidence for the construct validity of an ability test is obtained to the extent that good performance can be explained by examinees' following sound thinking processes, and to the extent that poor performance can be explained by deviations from such procesㅅses. Studıes of thinking processes can provide the information needed to judge the soundness of thinking processes used.

I. THE PROBLEM

The discussion thus far has argued that in principle the information gathered in studies of thinking processes ought to be relevant to construct validation. This is true only if the information on people's test thinking which these studies yield is an accurate reflection of the thinking which would have taken place had the people taken the test outside the study. Studies of thinking processes typically require that subjects provide introspective reports of the progress of their thinking, or provide reasons for their performance. It is not known whether such requirements alter thinking from what would have taken place under testing conditions in which such verbal reports are not provided.

The study addressed the following two general questions:

1. Are introspective reports of thinking reflective of the thinking that actually takes place? More specifically, does the accuracy of introspective reports of thinking depend upon the manner in which the report is elicited?

2. Do introspective reports of thinking reflect the thinking that takes place in testing situations in which only outcomes of thinking and not the thinking itself is reported? That is, does the elicitation of introspective reports change the course of thinking from what it would have been without the elicitation?

II. THEORETICAL PERSPECTIVE

Belief in the potential usefulness of studies of thinking processes is often motivated by the perspective of scientific realism. Fundamental to the philosophy of scientific realism is the view that scientific investigation is aimed towards, among other things, the identification of the underlying causes of directly observable phenomenon. The postulation of theoretical entities is taken to be speculation about the real constitution of the world, speculation which is then tested through further exploration.

Scientific realism is often contrasted with instrumentalism or positivism, views which do not concede the reality of theoretical entities. On these accounts theoretical entities are taken to be the imaginative speculations of scientists; useful fictions designed to bring coherence to a vast array of unconnected observables. To this view the scientific realist retorts: "But if theoretical entities are supposed to be the underlying causes of what is directly observed (a view which both instrumentalists and positivists usually espouse), how can they serve this role if they are fictions in the minds of scientists? Causes make things happen, a function which fictions in the minds of scientists are singularly unsuited to perform!"

The goal of construct validation is to discover the causes of performance on tests. When the tests are mental ability tests the attempt in test design is to make a test such that the designated mental ability is the cause of test performance, and construct validation is conducted to determine the extent to which this has been achieved. Mental abilities are assumed to underlie performance in the sense that they are not currently directly observable, but must be inferred from what can be observed. The observation of performance alone typically leads to highly ambiguous inferences about underlying abilities because competing possible causes of performance cannot be ruled out. Some method is needed to push back the bounds of the observable beyond typical sorts of test performance, a task for which studies of thinking processes are particularly well suited, and a task which must be accomplished for science to proceed (Norris, forthcoming).

The procedure is somewhat complicated by the fact that we are not able to specify the nature of mental abilities in advance of doing the scientific investigation. It is precisely this knowledge which the investigation is designed to achieve. Imagine, then, wanting to construct a test of mental ability "X". Not knowing the nature of ability X, how is it to be recognized that X and not some other ability the cause of test performance? At the current stage of the science of mental abilities, ability X is likely defined in terms of directly observable performances. Therefore, to take the performances alone as evidence that X is the operative cause of those performances is to reason in a circle. The issue is complicated, and cannot be resolved without the interplay of scientific and philosophical reasoning (Norris, 1985). Two pertinent questions which must be answered include: (i) Can we imagine *how* the operation of the postulated mental processes would produce the performances? and (ii) Are we *willing* to conclude that these processes are manifestations of the ability we are trying to test? There are no fixed rules for answering such questions, but it is clear they require deep thought by those thoroughly immersed in the field.

III. HISTORY OF STUDIES OF PROCESS

Studies of thinking processes have experienced a long history of endorsement by test validation theorists. For example, Cronbach (1971, p. 474) suggested that such studies can usually amplify the meaning of constructs. This endorsement is contrasted with very few reported examples of research of this type.

One of the earliest and most extensive studies in this tradition was conducted by B.S. Bloom and L.J. Broder in 1950 on the thinking processes of college students solving certain test problems. Bloom and Broder believed that inferences from test behaviour to underlying mental processes are untrustworthy unless they rely on explicit exploration of those processes. They knew that many mental processes could lead to the same performance, and provided examples of sound thinking leading to incorrect preformance and unsound thinking resulting in correct solutions. Their approach to gaining more direct information on mental processes was to have examinees think aloud while answering questions on a test. They found that by first giving practice on thinking aloud while solving some simple multiplication problems subjects were able to provide more detailed reports. One of their major conclusions was that "the method of thinking aloud served . . . to yield relatively consistent and meaningful data from the majority of subjects" (1950, p. 90).

Bloom and Broder failed to question the meaningfulness of the information contained in subjects' reports for situations in which the subjacts would not have been asked to think aloud. Just because the introspective reports were consistent and meaningful does not mean that they were reflective of thinking that would have taken place in other sorts of situations. If requiring people to think aloud as they work through test questions makes their thinking substantially different from what it would have been had they not thought aloud, then the information gathered in the validation study is not relevant to testing situations in which verbal reports of thinking are not sought.

In another study R.P. Kropp (1956) examined the relationship between thinking processes revealed in oral problem solving and the solutions provided to the problems. Like Bloom and Broder, Kropp concluded that verbal reports of thinking reveal a great variety of mental processes leading to the same answer. Kropp also concluded that the technique is useful for exposing ambiguities and hidden cues in test items. Still, the question of what is learned from think aloud contexts about normal test taking contexts, and the question of the accuracy of think aloud reports were no better understood after this study.

C. McGuire (1963) reported on an attempt to help improve the construction and interpretation of an examination by using experts' introspective judgements of the mental processes required to answer questions on it, and students' reports of the processes they followed while taking the test. She found that the method had a fair degree of usefulness in designing tests of more complex mental processes and in bringing student assessment into better agreement with the objectives of the instruction. Without further elaboration she also remarked that it became apparent that "the interview [technique] did not sufficiently simulate an examination situation to allow sound conclusions to be drawn" (p. 9). One cannot be certain, but I assume she meant sound conclusions about whether the results were applicable to situations in which introspective reports were not gathered. If this is what she intended it is a puzzling and disappointing fact that she did not explain her position further. At the same time, it is important that she recognized a problem which still needed to be explored.

In 1964 J.A. Connolly and M.J. Wantman reported a study which they considered to be an improvement upon the original one in this tradition by Bloom and Broder. Like Bloom and Broder, they assumed that inferences about the nature of reasoning processes drawn from typical item analysis statistics are tenuous at best (p. 59). In the study, subjects (of which there were only 9) were told to think aloud, reporting all thoughts that might cross their minds during their attempts to respond to a set of test items. No probing other that this non-directive instruction was used. As in the Bloom and Broder study, instances were found of good thinking coupled with incorrect answers and poor thinking with correct answers. Adequacy of thinking was rated in accord with a model of quality thinking on the test items. The overall conclusion was that the technique is useful for pretesting items in the construction of a test.

H. Schuman (1966) used the technique of probing people's reasons for the answers they chose on a test. The probing was conducted after the test was completed, with each individual probed on a randomly selected set of items from all those contained on the test. Responsos were evaluated on a five point scale, with a score of "1" given for an explanation which was quite clear and led to accurate prediction of the answer chosen, to "5" for an explanation that was very unclear and did not support any prediction about the answer chosen. Total scores for individuals over all items on which they were probed were calculated, the lower the score indicating the higher the individual's understanding of an item. Total item scores for each item over all individuals who were probed on it were also calculated, and indicated the group's understanding of the individual items. The qualitative information contained in the analysis of the verbal reports was also used to help understand "more precisely what [the analyst] is measuring -- which is, after all, the final goal of 'validity'."

A colleague and I used think aloud protocols in the development of a critical thinking test on appraising observations (Norris and King, 1984). Our desire was to conduct the interviews in a fundamentally nonleading fashion. We wished to influence students' thinking as little as possible, realizing that just asking them to think aloud and placing them alone with a stranger might have effects in themselves. Still, it seemed on occasion that interrupting a student's narrative might be more beneficial than not, particularly when the interruption was merely to clarify the ambiguous referent of a pronoun, or to point out obvious reading errors. Although we did not wish to rush examinees, to cut off reasoning by inadvertent signals, or to endorse or criticize particular reasoning attempts, we did wish to obtain records of reasoning which were as complete as possible. To fulfill this aim it was often necessary to probe beyond the initial instruction to think aloud. This probing was done only after examinees had chosen their answers to questions and had finished reporting on their thinking. Even in these follow-up stages probing was as nonleading as possible, merely echoing already reported thoughts or asking to explain choices of answers a little more fully. It was in the context of developing this test that the present study was conceived.

IV. RELEVANT STUDIES IN NON-TESTING CONTEXTS

The essential nature of studies of thinking processes is that they are attempts to extract information from people's memories, usually their short term or very recent memories. This fact suggests that research on creation of and extraction of information from memory would be relevant. Much of this research can be found in studies of the use of verbal reports as data in the information processing tradition and in studies of eyewitness testimony.

Verbal Reports as Data

Much of the work on the trustworthiness of verbal reports of mental processes is reviewed in one of three recent articles and a recent book (Ericsson and Simon, 1980, 1984; Nisbett and Wilson, 1977; Smith and Miller, 1978). The essence of the Nisbett and Wilson report is that people have little or no introspective access to the things which stimulate their cognitive processes. Ericsson and Simon and Smith and Miller are critical of this conclusion, and claim that people do have dependable access to their mental processes in certain situations.

Nisbett and Wilson conclude three things: (i) people often cannot accurately report the effects of certain stimuli on their responses to problems requiring higher order thinking; (ii) 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 (iii) when people are correct about the stimuli affecting their responses they have coincidentally employed a hypothesis which happens to be correct.

Not all of the studies reviewed by Nisbett and Wilson can be described here for the number is quite large. They relied on evidence from the cognitive dissonance literature, the self-perception attribution literature, the learning without awareness literature, and the literature on problem solving, among other fields. From studies on cognitive dissonance and self-perception they concluded that people can change their attitudes without any apparent awareness of such change, and can be motivated by things of which they are not aware. They argued that results from studies of problem solving suggest that experimental subjects are usually not aware of experimentally manipulated factors which have influenced their responses. They also review a series of studies designed to demonstrate people's inability to report accurately on the effects of experimentally controlled stimuli on their responses. For example, people are not aware of the effect that position on the rack has on their selection of a garment; are not aware of the effect of people's personalities on their assessment of those people's appearance; are unable to accurately report on the effect of distractions on their reactions to such things as a film; and are unable to accurately rate the effect of being assured there was no danger on their willingness to subject themselves to such things as electric shock.

Nisbett and Wilson do suggest situations in which accurate verbal reports can be expected. These are characterized by an available influential stimulus, a stimulus which is a plausible cause of the response, and a lack of other plausible causes of the response. The experimental situations upon which they base their conclusions do not meet all of these conditions. In particular, experiments are situations in which the influential stimulus is not available to the subjects because it is "systematically and effectively [hidden] from them by [the] experimental designs" (Smith and Miller, 1978, p. 356). It is on this point that Smith and Miller criticize Nisbett's and Wilson's conclusions most severely, because they apply only to situations (experimentally controlled ones) in which the outcome, subjects' unawareness of what was influencing their thinking, is what would naturally be expected. Nisbett's and Wilson's analysis does not inform us of whether in other situations people's mental processes are more accessible to them.

Another limitation of the Nisbett and Wilson analysis arises from the depth of the mental processes which they examined. They are dealing with subtle mental processes such as those which govern the formation of attitudes, which stimulate insightful solutions, and which bias evaluations. What kinds of mental processes are these? Are they of the sort that people can be aware of them? If they are the sort of process that, say, governs such things as the human heart beat or regulates breathing, then it is not surprising that people cannot access them. As Smith and Miller point out, people are not even able to report on the mental processes involved in less deep, but yet routine, processing such as that involved in producing answers to well-learned multiplication tables.

Ericsson and Simon (1980, 1984) discuss the trustworthiness of verbal reports on mental processes in light of a theory of thinking conceived as information processing. They conclude that instructions to verbalize do not change the course of cognitive processing, but merely slow it down, 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, since it is the provision of reasons for answers that is often sought. The conclusion suggests that information about test validity gathered in interview contexts might not be applicable to testing contexts in which interviewing was not done. With regard to the completeness of verbal reports of thinking, Ericsson and Simon conclude that certain types of things tend to be omitted. Processes that are so well learned that they have become automatic tend not to be reported, and often subjects are able to behave in accord with rules without being able to verbalize them.

One of the values of the Ericsson and Simon work is that it gives specific information on the situations in which one can expect verbal reports to be trustworthy, and on the ones in which they are justifiably mistrusted. In particular, their research indicates 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. However, it is not legitimate to assume that this research answers all the questions for testing situations. Testing contexts are sufficiently different from the ones in which information processing research is conducted that it is reasonable to assume that memory retrieval and information processing demands might also differ. In particular, taking tests is a situation that carries with it certain assumptions about how one should try to perform, how the results reflect upon the individual, and so on. These assumptions are probably different from those that go along with being involved in a study in which tests are not given, and possibly lead to different influ'nces on performance.

Eyewitness Testimony

Eyewitness testimony is often contained in verbal reports of what people can remember, or claim to be able to remember. These reports are often given in response to instructions of one sort or another. Reports of what examinees are thinking when responding to a test are similar sorts of things. In one situation people search their memories for recollections of what they have observed; in the other for what they have thought. It is reasonable to believe that the recall processes in both situations are related. Thus the eyewitness testimony literature, which contains information on the factors which affect the accuracy of reports of observations, is pertinent to the question of the accuracy of reports of thinking while taking tests. The degree of pertinence is tempered by the dissimilarities between the two situations: in one, recall of the recognition of an external event takes place, whereas in the other recall of an internal event occurs; in one, memory is probed about events in the more distant past, whereas in the other the memory is of events in the very recent past.

The most relevant eyewitness testimony research for the present study concerns the effect of different types of questioning on the accuracy of reports. Three categories of questions have been studied: (i) those eliciting _free_ reports (for example, "Tell us all that you saw"); (ii) those eliciting _controlled_ reports (for example, "Give us a description of what your assailant was wearing"); and (iii) those eliciting _alternate-choice_ reports (for example, "Did your attacker have dark or light hair?") (Loftus, 1979, p. 90). Two general conclusions can be drawn on the basis of many independent tests of the influence of these types of questioning techniques. The first 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 somewhat more, and alternate-choice reports the most 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. Independent support for these results has been given by many investigators including Clifford and Scott (1978), Dale, Loftus and Rathbun (1978), Harris (1973), Hilgard and Loftus (1979), Lipton (1977), Loftus and Palmer (1974), and Marquis, Marshall and Oskamp (1972). The results are also largely consistent with the theory and evidence offered by Ericsson and Simon.

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 the testing situation. Eliciting reports of thinking on tests is different from eliciting recollections of observed events, and there is no research which explores the relevance of these differences to factors affecting the accuracy of both types of report. In addition, testing is a different social context from involvement in psychological experiments, and it is not known how this fact would influence recall from memory.

V. METHOD

Sample

Five senior high schools were chosen on the east coast of Newfoundland, Canada. The communities in which the schools were located ranged from one-industry communities with less than 1000 people to a somewhat larger town of about 5000. The total sample consisted of 343 students which included all of the students in grades 10, 11, and 12 in four of the schools, and about half of those in the other. This sample provided for a broad range of student abilities. In addition, although all the schools were in small communities, they were within commuting distance of the capital city and indeed many of the teachers commuted every day. Thus, the schools experienced little trouble in attracting highly qualified teachers. In addition, the students in these schools scored at or above the national average on standardized measures of achievement.

Procedure

A completely randomized factorial design was used to study the effect of various levels of probing on examinees' thinking processes while they worked through Part A of the _Test on Appraising Observations_ (Norris and King, 1983). Four levels of probe were used: (i) _Think Aloud_, in which examinees were asked to report all they were thinking as they worked through the items; (ii) _Immediate Recall_, in which examinees were asked to tell why they had chosen the answer they did; (iii) _Criteria Probe_, in which a feature of each test item was mentioned and examinees were asked whether those features made any difference to the answers they chose; and (iv) *Principle Probe*, which was a criteria probe with the additional question of whether choices of answer were based upon particular general principles. The probes vary in degree of "leadingness" (according to standard concepts of what it is to be a leading question), and also vary in the task required. The first level of probe gives considerable leeway for examinees to report as they see fit, while the subsequent ones ask for particular sorts of information and are thus more directive of the task to be carried out.

An associate and I each selected students according to the order they appeared on class lists. They were taken from their classes one at a time and randomly assigned to one of the experimental groups, either one of the probe groups or a control group. Students falling into the probe groups were asked to first work through items 1-15 while they were interviewed. As they worked through each question they were asked to mark their answers on the answer sheet and either to think aloud, or to tell why they had chosen their answer, or to respond to either the criteria or principle probe. The reports were tape recorded. The remaining 13 items on Part A were then completed by the students working privately in a more normal testing situation. Students in the control group were not interviewed, and were asked to work privately through all 28 items on Part A while marking their answers on the answer sheet. The raw data thus consisted of answers marked on the answer sheets and tape recorded protocols for those in the experimental groups.

Data Analysis

Two sets of scores were derived from the raw data. The first set consisted of performance scores, numbers of items right according to the key provided with the test (Norris and King, 1985). Total number of questions correct was calculated for items 1-15 and for items 16-28. The second set of data consisted of thinking scores. Thinking scores were determined for items 1-15 for all students in the experimental groups. Scores reflected the quality of thinking displayed in the protocols on a scale of 0-3 for each item (Norris and King, 1984). Total thinking scores for items 1-15 were calculated.

The following three questions were addressed in a series of quantitative and qualitative analyses:

1. uo verbal reports of thinking on tests accurately portray thinking that takes place?

2. Is thinking concurrent with reporting different from thinking in testing situations in which reports of thinking are not elicited?

3. Is thinking subsequent to reporting different from thinking subsequent to testing situations in which reports of thinking are not elicited?

Quantitative Analyses

Question 1: Do verbal reports of thinking on tests accurately portray thinking that takes place? Verbal reports of thinking can be useful in the validation of ability tests which are to be used in situations where such reports are not elicited only when the reporting does not shift thinking from the course it would have followed had the reporting not taken place. However, even if this condition is satisfied, a further issue remains. Do verbal reports of thinking give an accurate portrayal of the course thinking follows, regardless of whether giving the reports changes the course of thinking? This is the issue addressed by question 1, and is the issue raised by the first general question posed at the outset of this report.

Trying to answer this question raises a vexing issue, for which only a compromise solution is currently available. The issue involves the availability of a criterion of accuracy of reports of thinking. In some sort of ideal situation, what the scientist would like to do is follow the course of thinking independently of the person engaging in it, by having a "window into the brain" or some such access. Then the match between the person's verbal reports of thinking and the scientist's independent observation of it could be compared and we would obtain a measure of accuracy of the verbal reports. No such ideal situation can currently be created, nor even approximated. In addition, while recognizing that a complex of interconnected experiments might provide inferential access to people's thinking, the source of information upon which the scientist must rely most extensively is the thinker's own verbal reports of thinking. But it is the accuracy of such verbal reports as portrayals of the course of thinking that is at issue in this study. Some compromise must be sought.

Indicators of accuracy have been suggested from time to time. Ericsson and Simon (1980) suggest that the investigator's judgement of the completeness of a subject's reasoning can indicate accuracy. If the reported reasoning la:ks something the investigator has good reason to believe was needed, then the report can be judged incomplete to this extent. While useful, this criterion depends on the imagination and insight of the investigator and for this reason is likely to be applied unevenly across situations. Schuman (1966) recommends gauging the accuracy of verbal reports by the extent to which they lead to correct predictions of the subject's choice of answer. To the extent that correct predictions can be made, the reports are judged accurate. One problem with this approach is that there are factors other than thinking which affect subjects' choices of answers. Thus, even perfectly accurate reports of thinking would not necessarily lead to accurate predictions of responses. In addition, sometimes accurate predictions can be made independently of any knowledge of subjects' thinking.

Given no clear best way to proceed, I decided to take the thinking scores obtained by examinees in the Think Aloud group to be the criterion against which to compare reports from the other groups. This approach assumes that differences in thinking processes among the groups would show up in differences in thinking scores, and assumes rather than studies the accuracy of the Think Aloud reports. However, it is generally conceded (Ericsson and Simon 1980, 1984; Loftus, 1979) that free reports such as those given in the Think Aloud group are the most accurate of all. The main issue concerns not the accuracy of what is reported, but the completeness. That which is reported is generally assumed to be trustworthy, but it is also assumed that aspects of thinking are not reported in such situations.

With Thinking Score as the dependent variable, and the Think Aloud group taken as the control, a 4 x 3 x 2 x 2 fixed effects analysis of variance was performed using the SPSS MANOVA procedure and with Interview. Group, Grade Level, Interviewer, and Sex as independent variables. This analysis allowed between 5 and 6 observations per cell given the 271 subjects in these four interview groups. The Non-interview group was excluded from this analysis.

Question 2: Is thinking concurrent with reporting different from thinking in testing situations in which reports of thinking are not elicited? If it can be concluded that verbal reports provide accurate portrayals of thinking that is taking place, the first condition for the usefulness of studies of thinking processes to test validation has been met. The second condition, raised by the second general question at the beginning of this report, requires that eliciting the verbal reports does not itself affect the course of thinking. If it does, then the usefulness of studies of process would be limited to testing situations in which verbal reports of thinking are also elicited. Such types of tests are possible, and maybe even desirable. But given the time required for their administration, and the attendant costs, they are not likely to achieve wide use.

If eliciting thinking reports alters the course of thinking, then this should be manifested in different performances between those being interviewed and those taking the test without being interviewed. With Total Performance Score on items 1-15 as the dependent variable, and the No Probe group as the control, a 5 x 3 x 2 x 2 fixed effects analysis of variance was performed with Interview Group, Grade Level, Interviewer, and Sex as the independent variables. This allowed between 5 and 6 observations per cell using the total sample of 343 subjects.

Question 3: Is thinking subsequent to reporting different from thinking in testing situations in which reports of thinking are not elicited? It is widely believed that in addition to illustrating what they know, people often acquire new knowledge while taking tests. This fact needs to be taken into account in the interpetation of test scores, although knowledge is not yet sufficient for doing this well. However, if in presenting verbal reports of thinking examinees learn different things from when they do not provide such reports, then the usefulness of studies in the former context is diminished for the validation of tests used in the latter context. Question 3 thus also addresses the issue raised by the second general question.

With Total Performance Scores on items 16-28 as the dependent variable and the No Probe group as the control, a 5 x 3 x 2 x 2 fixed effects analysis of variance was performed using the same independent variables used in the analyses of Question 2. In order to simplify interpretations, in the analyses for all three questions the four-way interaction mean square was combined with the error term.

Qualitative Analyses

Quantification often entails the loss of some "richness" of information. In particular, representing protocols by a series of thinking scores as was done in this experiment is bound to lose some of the information contained in the original verbal reports. As an alternative approach to answering Question 1, I conducted a qualitative analysis of a random sample of 40 (stratified by treatment group) of the total sample of 271 interviews. The following seven categories of verbal moves for describing the protocols resulted from this analysis: (i) Reference to Details - either recalling a factual detail given in an item prior to one currently being worked on, recalling such a prior detail incorrectly, or stating a detail in the current item; (ii) Asking Rhetorical Questions - posing questions which appeared to be directed to the examinee himself or herself rather than to the interviewer; (iii) Making Self-Evaluations - either evaluating judgements or conclusions which had been previously explicitly stated, or evaluating ones which had not been verbalized; (iv) 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; (v) Using Attention Control Devices - either making comments about the stage of progress reached in reasoning through the problem (Let's see, Where was I, etc.), or commenting on the direction reasoning should proceed (Wait now); (vi) Interacting with the Experimenter - directing comments or questions to the experimenter; and (vii) Pausing - either making verbal inflections (Ahhh, Mmmm, etc.) or being silent.

Protocols were coded according to the seven categories and occurrences were accumulated for each category across the forty subjects. No sophisticated statistical analysis was performed. At this stage the data were taken to be purely exploratory, and were examined merely for general trends with a view to more systematic exploration in the future. The question asked was whether interview group membership affected the course of thinking in ways that were not detectable by differences in thinking scores, but were detectable by the above seven categories.

VI. RESULTS

Question 1

Table IA gives the main effect means for each level of the four factors examined. An examination of the table indicates differencec on the order of 1 point or less. Table IB gives the analysis of variance summary for the four factors. The analysis revealed no significant interaction or main effects. The Interview Group main effect was nonsignificant.

The qualitative analysis of the 40 randomly chosen protocols also revealed little differnnce among interview groups, which was the factor of primary concern in this study. Given the qualitative and speculative nature of the seven cıtegories which were developed to describe the protocols, no sophisticated statistical analysis of the data was performed. Rather, the results were examined for obvious trends which would indicate some interesting differences to explore more rigorously. No such differences were found. Table IC is a contingency table of the seven categories against interview group. While there are clear differences between the protocol categories, with some having occurrences on the order of hundreds of times and others on the order of tens of times, there are no glaring differences in trend between interview groups. The categories register occurrences with the same order of magnitude across all groups. It did not seem reasonable to try to pry more than this conclusion from this data.

Question 2

Table IIA contains the main e:fect means for each level of the four factors examined. Visual inspection of the table indicates that al! differences are small, being on the order of about 0.5 on the performance scale. Table IIB gives the analysis of variance summary for the four factors. The analysis showed no significant interaction effects, and a significant main effect for interviewer.

Question 3

Table IIIA contains the main effect means. Inspection shows that there are only very small differences for all factors. Table IIIB contains the analysis of variance summary infcrmation and shows significant effects for Interviewer, Sex and Grade Level. There are no significant interaction effects, and the effects for Interview Group are nonsignificant.

Table IA

Main Effect Means: Question 1
Accuracy of Verbal Reports

Factor	Level	Mean
Interview Group	Think Aloud	7.9
	Immediate Recall	9.2
	Criteria Probe	8.8
	Principle Probe	9.0
Interviewer	A	8.1
	B	9.3
Sex	Male	9.2
	Female	8.3
Grade	Level I	8.2
	Level II	8.6
	Level III	9.5

Table IB

Analysis of Variance Summary: Question 1

Source	df	MS	F
Main Effects			
Interview Group	3	13.8	0.868
Interviewer	1	26.0	1.64
Sex	1	32.4	2.04
Grade	2	42.9	2.70
Two Way Interactions	17	24.4	1.53
Three Way Interactions	17	18.1	1.14
Residual	229	15.9	

Table IC

Qualitative Analysis: Question 1

	Think Aloud	Immed. Recall	Crit. Probe	Princ. Probe
Reference to Details	104	139	99	139
Rhetorical Questions	16	9	2	5
Self-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

Table IIA

Main Effect Means: Question 2
Thinking Concurrent with Reporting

Factor	Level	Mean
Interview Group	No Probe (Control)	7.8
	Think Aloud	8.0
	Immediate Recall	8.3
	Criteria Probe	7.9
	Principle Probe	7.6
Interviewer		
	A	7.6
	B	8.2
Sex		
	Male	7.7
	Female	8.0
Grade		
	Level I	7.8
	Level II	7.7
	Level III	8.1

Table IIB

Analysis of Variance Summary: Question 2

Source	df	MS	F
Main Effects			
Interview Group	4	5.40	1.02
Interviewer	1	17.8	3.35*
Sex	1	3.70	0.695
Grade	2	4.56	0.857
Two Way Interactions	21	5.20	0.977
Three Way Interactions	22	4.75	0.893
Residual	290	5.32	

* $p<0.01$

Table IIIA

Main Effect Means: Question 3
Thinking Subsequent to Reporting

Factor	Level	Mean
Interview Group	No Probe (Control)	8.4
	Think Aloud	8.4
	Immediate Recall	8.3
	Criteria Probe	8.6
	Principle Probe	8.1
Interviewer	A	8.2
	B	8.5
Sex	Male	8.0
	Female	8.7
Grade	Level I	7.8
	Level II	8.6
	Level III	8.8

Table IIIB

Analysis of Variance Summary: Question 3

Source	df	MS	F
Main Effects			
Interview Group	4	1.93	0.429
Interviewer	1	12.9	2.88*
Sex	1	32.3	7.19**
Grade	2	34.5	7.70**
Two Way Interactions	21	6.43	1.44
Three Way Interactions	22	4.59	1.02
Residual	290	4.48	

* $p<0.05$
** $p<0.01$

VII. DISCUSSION

Question 1

Do verbal reports of thinking on tests accurately portray the thinking that takes place? The results of this study show that the accuracy of reports in portraying the essential elements of the thinking process on a critical thinking test does not vary across a variety of probing techniques, from the nonleading elicitation of free reports to the leading elicitation of controlled reports. There were no significant differences in the quality of thinking as measured by Thinking Scores across the four levels of probe studied. In addition, the qualitative analysis of protocols revealed that there was no essential difference in the verbal moves used in reporting under different elicitation procedures. Both results suggest strongly that it is subjects' thinking and not how that thinking is elicited that controls what is reported. If this result can be substantiated, then it would seem that the accuracy of verbal reports of thinking on tests is not as sensitive to the type of probing as research in other contexts would indicate.

The issue of the criterion of accuracy must always be kept in mind, though. There is no available technique, nor is there likely to be one in the near future, for gaining direct access to people's thinking processes independently of their introspective observations. To conduct this study, we assumed that the most accurate reports could be obtained from asking subjects to think aloud, with no further probes being made.

This would seem to provide the least amount of interference possible while still eliciting the desired information. When compared to this group, the other groups provided equally accurate reports. The question remains about the degree to which free reports are an accurate reflection of thinking that takes place. Accuracy in this context is a function of two considerations, whether as far as they go reports accurately describe the thinking process, and whether reports go far enough in giving complete descriptions of the entire thinking process. It is doubtful that verbal reports of thinking are ever fully complete, for there appears to be much thinking for which we have little or no introspective access (Nisbett and Wilson, 1977). There can be some confidence that what is reported is an accurate reflection of that aspect of the thinking process which is described. This study suggests that in testing contexts such as those employed, the degree to which a probe is leading does not affect the accuracy of thinking reports.

Question 2

Is thinking that occurs concurrent with reporting on thinking different from thinking in testing situations in which reports of thinking are not elicited? Regardless of the accuracy of verbal reports of thinking, if such reports are to be useful in the validation of ability tests in which reports of thinking are not sought, then eliciting them cannot alter the course of thinking. If the course of thinking is altered by having people report on their thinking, then this alteration could be revealed in altered performance. Thus, performances which are systematically similar provide evidence that thinking is similar, though of course there is no necessity that similar performance result from similar thinking.

The results showed that there are no significant differences in performance on items 1-15 of the Test on Appraising Observations between those who reported their thinking while working on those items and those who worked on them alone while giving no reports of their thinking. This was true for all levels of probing, suggesting that probing did not alter thinking. If this suggestive result can be substantiated, then there are implications for the usefulness of this technique that extend beyond test validation contexts. The technique should also prove useful for conducting basic research into the nature of human reasoning, a use which has already been endorsed strongly by Ericsson and Simon (1984).

Question 3

Does the eliciation of reports of thinking have any effect on thinking which occurs subsequent to the elicitation? Such longer term effects could occur even if there are no immediate effects. In this study very long term effects were not examined. Rather, effects on performance were studied immediately after the reports were made. When subjects finished reporting their thinking on items 1-15 they were asked to work on their own on the remaining items on Part A of the test, items 16-28. The results showed no significant performance differences between those students who had been probed on items 1-15 and those who had not been, again suggesting that no significant differences in thinking occurred. There seems, then, to be no effects carried over from the reporting sessions which are highly relevant to how students think and perform on similar tasks immediately thereafter.

Type II Error

Whenever failure to reject the null hypothesis is the desirable result of an experiment, the power of the test to reject a false null hypothesis becomes an overriding concern. Was this experiment sufficiently powerful to detect any true differences which existed among the treatment groups? There are a number of reasons which make it highly plausible to believe that differences would have been detected had they been present in the population. The first turns on the fact that the treatments were considerably different from one another. It is quite a different situation for high school students to work alone on a test in a fashion they are well used to in school, from their working in the presence of a stranger who is probing their thinking in a way that hardly ever happens in school. Thus, if elicitations of thinking have an effect on the course of thinking, then it should have been revealed in differences in performance between the interviewed and uninterviewed groups. In addition, the interview treatments themselves were highly different. The leading probes were quite leading in that 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. Instead, students would regularly deny that the suggested factor had anything to do with their thinking and proceed to explain how their choices were made.

Another reason making the null results of this experiment plausible is that effects were sought from a number of different directions, but none were found in any of them. The quantitative analyses showed that no differences were detected either in the ratings of students' thinking or in ratings of their performance both during and after the interview sessions. In addition, the qualitative analysis showed that the same patterns of verbal moves were used by each treatment group. It is plausible to think that if differences existed they would have been detected by at least one of these methods.

In addition, it must be noted that psychological research uncovers consistent effects using similar sorts of treatments in studies of eyewitness testimony. This does not mean that differences should have been found in this study, but it does mean that <u>if</u> differences existed they should have been detected. Of course the demand for an explanation for why no differences exist in the situation studied in this experiment arises at this point. Although it is highly tentative at this time, there is evidence which suggests that the results of psychological research on the evaluation of eyewitness testimony are not always substantiated in studies of the practice of juries in actual courtroom situations. For example, although psychological research conducted in laboratory contexts suggests that "jurors" place an unwarranted amount of confidence in eyewitness testimony, studies of real jurors show no such tendencies (McCloskey and Egeth, 1983). One possible explanation of this fact is that the gravity of the situation induces jurors to realize that being sceptical of evidence is important to maintaining the presumption of innocence of the accused. No such importance is attached to psychological experiments.

It is possible that a similar sort of mechanism might have operated in the context of this experiment. The study required students to take a test, and in our society tests are typically treated seriously. Even when it is known that the results will have no long-term consequences for school grades or any such matter, it is highly probable that they will still be taken seriously. Not many students are likely to portray themselves as being less capable than they actually are by deliberately performing poorly. At least it has been my experience that students take very seriously the situations I present them. It is possible that this fact creates a certain resistance to being led by suggestive questions which resulted in the null result of the experiment.

In addition to these considerations, an analysis of the statistical power of the experiment was performed using techniques described in Kirk (1968, pp. 107-108). The analysis requires the calculation of a parameter and the use of charts based upon a procedure by Tang (1938) which require the setting of a probability of Type I error and knowing the degrees of freedom for the treatment and error effects. The parameter is given by:

$$\phi = \frac{\sqrt{\sum_{j=1}^{k} \beta_j^2 / k}}{\sigma_\epsilon / \sqrt{n}}$$

where:

$\sum_{j=1}^{k} \beta_j^2$ = sum of squared treatment effects

n = size of the jth sample

σ_ϵ^2 = error variance.

For the purposes of the calculation, $(k-1/n)(MS_{BG} - MS_{WG})$ was taken as an unbiased estimate of the sum of squared treatment effects, and MS_{WG} as an unbiased estimate of the population error variance. With the probability of a Type I error set at 0.05 for each analysis, the results showed that the power of rejecting the null hypothesis when it was false was >0.97 for Question 1, >0.96 for Question 2, and >0.99 for Question 3. These results coupled with the earlier considerations make the null result of this experiment highly plausible.

VIII. CONCLUSION

This research points to a useful validation technique for the testing field. Studies of thinking processes using the verbal reports of examinees has always seemed an obvious way to gather data on the construct validity of tests. Such studies have long been known to be time consuming and expensive, but also their usefulness and justifiability has been uncertain. This study provides examples in the context of validating a critical thinking test of how such studies of process might be conducted using different questioning procedures. The results of the experiment indicate that the researcher did not have to be overly cautious about the "leadingness" of the questions used to elicit reports of thinking. Basically, examinees appeared not to be easily led when reporting on their thinking. Comparisons of the quality of thinking displayed in the verbal reports, of overall performance on the test, and of the verbal moves made while reporting showed no significant differences from one interview group to another. In addition, performance scores for the interview groups did not differ significantly from the noninterviewed control group.

It is not known whether results similar to these would be found with all types of test items or with all types of content. Given the lack of knowledge in this area, prudence would suggest repeating this experiment for tests with other item types and in other content areas. If such were to be done, then the research reported here can serve as a prototype for these subsequent studies.

REFERENCES

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

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

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

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

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

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

Ericsson, K.A. and Simon, H.A. Verbal reports as data. Psychological Review, 1980, 87, 215-251.

----------------------. Photocol analysis: Verbal reports as data. Cambridge, MA: MIT Press, 1984.

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

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

Kirk, R.E. Experimental design: Procedures for the behavioral sciences. Belmont, CA.: Wadsworth, 1968.

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

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

Loftus, E.F. Eyewitness testimony. Cambridge, Mass.: Harvard University Press, 1979.

Loftus, E.F. and Palmer, J.C. Reconstruction of automobile destruction: An example of the interaction between language and memory. Journal of Verbal Learning and Verbal Behavior, 1974, 13, 585-589.

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

McCloskey, M. and Egeth, H.E. Eyewitness identification: What can a psychologist tell a jury? American Psychologist, 1983, 38, 550-563.

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

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

Norris, S.P. The choice of standard conditions in defining critical thinking competence. Educational Theory, 1985, 35, 97-107.

---------------------. The philosophical basis of observation in science and science education. Journal of Research in Science Teaching, forthcoming.

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

---------------------. The design of a critical thinking test on appraising observations. St. John's, Newfoundland: Institute for Educational Research and Development, Memorial University of Newfoundland, 1984.

---------------------. Test on appraising observations manual. St. John's, Newfoundland: Institute for Educational Research and Development, Memorial University of Newfoundland, 1985.

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

Saith, E.R. and Miller, F.D. Liaits on perception of cognitive processes: A reply to Nisbett and Wilson. Psychological Review, 1978, 85, 355-362.

Tang, P.C. The power function of the analysis of variance tests with tables and illustrations of their use. Statistics Research Meaorandum, 1938, 2, 126-149.