DOCUMENT RESUME

ED 302 825 CS 009 476

AUTHOR Norris, Stephen P.
TITLE Verbal Reports of Thinking as Data for Validating Multiple-Choice Tests. Technical Report No. 445.
INSTITUTION Bolt, Beranek and Newman, Inc., Cambridge, Mass.; Illinois Univ., Urbana. Center for the Study of Reading.
SPONS AGENCY Office of Educational Research and Improvement (ED), Washington, DC.; Social Sciences and Humanities Research Council of Canada, Ottawa (Ontario).
PUB DATE Jan 89
CONTRACT OEG-0087-C1001
GRANT 418-81-0781
NOTE 22p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Cognitive Processes; Critical Thinking; Grade 12; High Schools; *Multiple Choice Tests; Protocol Analysis; Reading Research; *Test Validity
IDENTIFIERS *Verbal Reports

ABSTRACT

A study examined the effect of verbal reporting of students' thinking on their performance during an examination. Subjects, 343 high school seniors, were randomly divided into 4 experimental groups, and a different procedure for eliciting students' thinking during a critical thinking test was used for each group. A control group took the same test in paper-and-pencil format. Results indicated that there were no significant differences in either test performance or quality of thinking among the five groups. The results indicated that verbal reports of thinking do not influence students' thinking and performance during exams, making them a potentially useful source of validation information. (Five tables of data are included, and 38 references are attached.) (Author/RS)

CENTER FOR THE STUDY OF READING

A READING RESEARCH AND EDUCATION CENTER REPORT

Technical Report No. 445

VERBAL REPORTS OF THINKING AS DATA FOR VALIDATING MULTIPLE-CHOICE TESTS

Stephen P. Norris
Institute for Educational Research and Development
Memorial University of Newfoundland
and
Center for the Study of Reading
University of Illinois at Urbana-Champaign

January 1989

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

The work upon which this publication was based was supported by the Social Sciences and Humanities Research Council of Canada (Grant No. 418-81-0781). 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.

Abstract

Verbal reports of examinees' thinking on test items can provide useful validation data only if the verbal reporting does not change the course of examinees' thinking and performance. Using a completely randomized factorial design, 343 senior high school students were divided into five groups. In four of the groups, different procedures were used to elicit students' thinking as they worked through Part A of a critical thinking test of observation appraisal (Norris & King, 1983). In the control group, students took the same test in paper-and-pencil format. There were no significant differences in test performance among the five groups, nor in the quality of thinking among the four elicitation groups. These results are evidence that verbal reports of thinking meet one of the necessary conditions of useful validation data, namely, that collecting the data not influence examinees' thinking and performance. Since verbal reports of thinking can also contain a wealth of information on the psychological processes that underlie performance, they are a potentially important source of validation information.

VERBAL REPORTS OF THINKING AS DATA FOR VALIDATING MULTIPLE-CHOICE TESTS

Verbal reports of examinees' thinking are often recommended as relevant and important sources of evidence for validating tests (Anastasi, 1988; Cronbach, 1971; Ennis & Norris, in press; Haney & Scott, 1987; Messick, in press; Norris, in press-b, in press-c). Sometimes the proposed relevance is indirect, as when verbal reports of thinking are used to develop information processing models of test performance which, in turn, are directly relevant to assessing construct validity (Embretson, 1983; Embretson, Schneider, & Roth, 1986). Verbal reports of thinking have been used in test validation (Bloom & Broder, 1950; Connolly & Wantman, 1964; Haney & Scott, 1987; Kropp, 1956; McGuire, 1963; Schuman, 1966) but, possibly because of past emphasis on behavioristic approaches, not extensively. With the growing emphasis on cognitive approaches, it is likely they will receive greater attention (Afflerbach & Johnston, 1984; Ericsson & Simon, 1980, 1984), so studies of their usefulness for validating tests that go beyond mere recommendations and theoretical rationales are needed.

This study examined the relevance of the data from verbal reports of thinking on test items for validating multiple-choice tests that would be taken normally in paper-and-pencil format. A necessary condition for the data to be relevant is that the verbal reporting not alter examinees' thinking and performance from what it would have been had they taken the test in its paper-and-pencil format. The satisfaction of this condition is often taken for granted, but this assumption is not warranted. There is no firm evidence which shows whether or not asking examinees to report on their thinking while taking tests affects the course of their thought. The purpose of the study was to gather such evidence.

There is pertinent evidence on the effects of verbal reporting on the course of thought from outside testing contexts. For example, research on eyewitness testimony has shown that testimony given in response to non-leading questions tends to be more accurate than testimony given in response to leading questions (Clifford & Scott, 1978; Dale, Loftus, & Rathbun, 1978; Harris, 1973; Hilgard & Loftus, 1979; Lipton, 1977; Loftus, 1979; Loftus & Palmer, 1974; Marquis, Marshall, & Oskamp, 1972). This result is pertinent to the problem raised here to the extent that the mental processes used to report eyewitness testimony are the same as those used to report one's thinking on test items. Some of the processes are likely the same, since both activities involve memory retrieval. But not all the processes are likely the same: In the eyewitness testimony situation there is recall of an observation of an external event whereas, in the testing situation, there is recall of an internal thinking process; in the eyewitness testimony situation memory is probed about events in the more distant past whereas, in the testing situation, memory is probed about events in the very recent past.

Evidence from research on information processing is also pertinent to determining the effect of verbal reporting on the course of thought. Ericsson and Simon (1980, 1984) have concluded that instructions to verbalize thinking do not change the course of that thinking, but merely slow it down, when subjects are verbalizing information that would be available normally in short-term memory. However, they claim that specific and directive probes, especially requests for motivations and reasons, alter cognitive processing. These findings are particularly important because, if they generalize to the testing context, they cast doubt on recommendations to use such validation techniques as "analysis of reasons" (Messick, in press), which probe for examinees' reasons for answers. However, it is not known whether or not they do generalize.

This study addressed the following general question: Does the elicitation of verbal reports of thinking on multiple-choice items requiring deliberative thought alter the course of examinees' thinking and performance on those items from what it would have been had they answered the items in paper-and-pencil format without reporting verbally on their thinking? Only if the answer is negative can verbal reports of thinking on multiple-choice tests requiring deliberative thought be relevant to the validity of those tests in the context of their paper-and-pencil use. However, even if the answer is negative, this does not automatically mean that verbal reports of thinking are useful for multiple-choice test

validation. Maybe, for instance, the verbal reporting does not alter the course of examinees' thinking and performance, but reveals so little about their thinking that it is worthless. The study did not directly address this issue, but nevertheless provided some information on it.

The focus of the study was the validation of multiple-choice tests that require deliberative thought. I am not concerned here with tests that require rote recall, but rather ones that require deliberate reasoning to figure out the answers. This is a broad and somewhat vague category. It includes tests of higher order thinking within specific school subjects, tests of critical thinking, tests of inference in reading, and problem solving and decision making tests. I focussed on multiple-choice tests for three reasons: (a) they are widely used because they fit very well the pragmatic constraints of many testing situations; (b) they are widely criticized as tests of deliberative thought (e.g., McPeck, 1981; Petrie, 1986) on the grounds that they provide weak evidence on thinking processes; and (c) it is this very weakness (if it exists) in the evidence that multiple-choice tests provide on thinking processes that verbal reports of thinking can plausibly eliminate.

Method

Sample

Five senior high schools were chosen from communities on the east coast of Newfoundland, Canada. The communities ranged from single-industry fishing and industrial communities with less than 1,000 people to a somewhat larger town of about 5,000, situated close to several similarly sized communities. This group of communities had a diverse economic base in fishing, government offices (including a police headquarters, a jail, and a court), tourism, light manufacturing, and shopping malls. The total sample consisted of 343 students, including 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 represented a broad range of student abilities. 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. The students in these schools scored at or above the national average on the Canadian Test of Basic Skills.

Instrumentation

The task was supplied by Part A of the Test on Appraising Observations (Norris & King, 1983). The Test on Appraising Observations is a multiple-choice test of one aspect of critical thinking, the ability to judge the credibility of reports of observations. The test has been rated highly in a recent survey of tests for assessing higher order thinking (Arter & Salmon, 1987). Part A has 28 items written in the context of a traffic accident at an intersection. In each item two people, either witnesses to the accident or individuals involved in it, report on what they observed happening. Examinees are to judge which, if either, of the reports is more credible. Relevant factors to consider in making judgments include the observer's expertise, alertness, and conflict of interest; the satisfactoriness of the observation conditions; and the source of the observation and the statement reporting it.

Here is Item 1 as an example:

> A policeman is questioning Pierre and Martine. They were in their car at the intersection but were not involved in the accident. Martine is the driver and Pierre, who had been trying to figure out which way to go, is the map reader.
>
> The policeman asks Martine how many cars were at the intersection when the accident occurred. She answers, "There were three cars."
>
> Pierre says, "No, there were five cars."

Examinees are instructed to choose which, if either, of the two underlined statements they have more reason to believe. The item is intended to test ability to recognize that the driver is likely to be more alert to the road conditions than the map reader and, therefore, that Martine's report is more credible, since all other factors appear equal.

Procedure

A completely randomized factorial design was used to study four ways of eliciting verbal reports of subjects' thinking as they worked on the test. Students were selected one at a time according to the order of alphabetical class lists. They were assigned randomly to one of five groups, either to one of four elicitation groups or to a control group. The groups are described in Table 1. An associate and I worked with students independently, each of us choosing the next available student on the list.

The verbal report elicitation procedures vary in the degree to which they lead examinees to provide particular sorts of information. The think aloud elicitation gives subjects the freedom to report as they see fit, and parallels the "free report" which yields the most accurate eyewitness testimony (Loftus, 1979). Subsequent elicitations request particular types of information and are thus more directive of the task to be carried out. The immediate recall elicitation requests reasons for answers selected, and was thus used to test the efficacy of Messick's (in press) proposed "analysis of reasons" and Ericsson's and Simon's (1980, 1984) claim that requests for reasons alter the course of thinking. The criteria probe and principle probe elicitations attempt to lead examinees by the questions asked, and thus were used to study the generalizability of the results from eyewitness testimony research on leading questions. In each group, subjects were told that they could go back to change their answers at any time. As an example, the elicitation procedures for Item 1 are described in Table 2.

[Insert Tables 1 and 2 about here.]

Tape-recorded verbal reports of thinking on items 1-15 were obtained from subjects in the elicitation groups. These subjects completed the remaining 13 items on Part A working privately in a paper-and-pencil format. Subjects in the control group worked privately in a paper-and-pencil format through all 28 items on Part A.

From the raw data, three sets of scores were derived. The *concurrent performance score* for each subject equalled the total number of items 1-15 answered correctly according to the key provided in the test manual (Norris & King, 1985). The *subsequent performance score* for each subject equalled the total answered correctly for items 16-28. The scores were called "concurrent" and "subsequent" because, for the elicitation groups, items 1-15 were done concurrently with verbal reporting and items 16-28 were done subsequently to it, working privately in a paper-and-pencil format.

A *thinking score* was assigned for items 1-15 for all subjects in the elicitation groups. For each item, the quality of each subject's critical thinking displayed in his or her verbal report was rated on a scale of 0-3 in accord with the procedure in Norris and King (1984) and these ratings totalled over the 15 items for each student. Thinking scores were assigned independently of the answers chosen.

Results

There were two main results: (a) the elicitation of verbal reports of thinking did not alter subjects' performance and, by inference, did not alter their thinking; and (b) the different procedures for eliciting verbal reports yielded essentially the same information of the quality of subjects' thinking.

Verbal Reporting and Performance

Two analyses support the conclusion that verbal reporting did not alter test performance. In the first, concurrent performance score was the dependent variable. This analysis determined whether giving

verbal reports of thinking affected ongoing performance. In the second analysis, subsequent performance score was the dependent variable. The analysis determined whether there was a carry-over effect from verbal reporting, possibly as a result of learning different things through the verbal reporting.

Two 5 x 3 x 2 x 2 fixed effects analyses of variance were performed with interview group. grade level, interviewer, and sex as the independent variables. This allowed on average between 5 and 6 observations per cell using the total sample of 343 subjects. In both analyses, the four-way interaction mean square was combined with the error term.

Table 3 contains mean concurrent and subsequent performance scores for each level of the four factors examined. All differences among means are small, being on the order of about 0.5. Neither analysis showed significant interaction effects. For concurrent performance, there was a significant main effect for interviewer only ($F(1,290) = 3.35$, $p < .05$). No significant differences in performance were found among the elicitation levels. For subsequent performance, significant differences for interviewer ($F(1,290) = 2.88$, $p < .05$), sex ($F(1,290) = 7.19$, $p < .01$), and grade level ($F(2,290) = 7.70$, $p < .01$) were found. Again, no significant differences were found among the elicitation levels.

[Insert Table 3 about here.]

Verbal Reporting and Quality of Thinking

Two analyses were performed: a quantitative and a qualitative. In the quantitative analysis, thinking score was taken as the dependent variable and elicitation group, grade level, interviewer, and sex as independent variables in a 4 x 3 x 2 x 2 fixed effects analysis of variance. This analysis allowed on average between 5 and 6 observations per cell given the 271 subjects in the four elicitation groups. The control group was excluded from this analysis, since they had not given verbal reports of their thinking and therefore could not be given thinking scores.

Table 4 gives mean thinking scores for each level of the four factors. Differences are on the order of 1 point or less. On the 15 item section, subjects averaged less than 1 point per item out of a total possible of 3 points per item.[1] No significant interaction or main effects were found.

[Insert Table 4 about here.]

A qualitative analysis of the course of students' thinking was conducted of a random sample of 40, 10 from each elicitation group, of the total sample of 271 students who gave verbal reports. Seven categories of verbal moves were derived from the verbal reports of thinking:

1. *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;

2. *Asking Rhetorical Questions* - posing questions which appear to be directed to the subject himself or herself rather than to the interviewer;

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

4. *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;

5. *Using Attention Control Devices* - either making comments about the stage of progress reached in reasoning through the problem, or commenting on the direction reasoning should proceed;

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

7. *Pausing* - either interjecting (Ahhh! Mmmm!), or being silent.

The verbal reports were coded according to the seven categories and occurrences were accumulated across the 10 subjects for each category. No statistical analysis was performed. The data were taken as exploratory and examined for general trends with a view to more systematic exploration in the future. The question asked was whether elicitation group membership affected the course of thinking in ways that were detectable by the above seven categories. The frequencies of each verbal move recorded in Table 5 suggest little systematic difference among elicitation groups. While there are clear differences among the verbal move categories, with some having occurred on the order of hundreds of times and others on the order of tens of times, a striking feature is that the order of magnitude of the frequency for each verbal move is the same for each elicitation group.

[Insert Table 5 about here.]

Discussion

The results support the conclusion that verbal reports of thinking on multiple-choice test items can provide relevant data on the validity of the tests taken in paper-and-pencil format. The conclusion has long been supported on theoretical and intuitive grounds. But it was not known whether a necessary condition for the relevance of verbal reports was satisfied, namely, that the reporting process not alter the course of thinking and performance from what it would have been had the test been taken in paper-and-pencil format. The results provide evidence that the condition is satisfied.

The analysis of verbal reporting and performance showed that test performance under a variety of elicitation procedures, from the nonleading request to think aloud to the leading questions about the role of specific pieces of information, is the same as performance in a paper-and-pencil sitting with no elicitation. The best explanation of this equivalent performance is that, on average, subjects in the elicitation and control groups thought equivalently. If eliciting the verbal reports altered the course of subjects' thinking, then this alteration should have been manifested in different performance scores between the elicitation groups and the control group. While theoretically possible, it is hard to imagine how subjects in the elicitation and control groups could have performed equivalently but thought significantly differently.

The analysis of verbal reporting and quality of thinking showed that there were no significant differences in the quality of thinking, as measured by thinking scores, across the four elicitation groups. The qualitative analysis of verbal reports revealed that there was no essential difference in the patterns of verbal moves used in reporting under different elicitation procedures. These results suggest strongly that it is the task presented by the items and not how subjects' thinking is elicited that governs what they report. Overall, the results support the use of verbal reports of thinking in validating multiple-choice tests.

Furthermore, the results suggest that special care need not be taken to avoid leading questions when eliciting reports of thinking, because examinees were not led easily when reporting on their thinking. Nevertheless, prudence may suggest a more cautious approach Given the evidence on the effect of leading questions in other domains and given that there was basically no difference in the information obtained using either elicitation procedure, it may be more sensible to use the least directive (think aloud) elicitation. A similar note of caution can be extended to Messick's (in press) proposal to analyze subjects' reasons for their answer choices as a source of data on validity. Given that Ericsson and Simon (1980, 1984) specifically caution that requests for reasons alter the course of thought and given that such requests seem to deliver nothing beyond a request to think aloud, the latter approach might be preferred.

Type II Error

Was this experiment sufficiently powerful to detect any true differences which existed among the groups? There are a number of reasons that make it highly plausible that differences would have been detected had they been present in the population. The first is the fact that the elicitation procedures were considerably different from the control procedure. It is quite different for high school students to work alone on a test in a way that normally occurs in school than to work in the presence of a stranger who is probing their thinking in a way that hardly ever happens in school. Thus, if elicitations of verbal reports of thinking have an effect on the course of performance, then it should have been revealed in differences in performance between the elicitation and control groups.

A second reason for thinking that any true differences would have been detected is that the elicitation procedures were considerably different from each other, but produced no differential effects. 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 easy for students to conform to these suggestions. Instead, they regularly denied that the suggested factor had anything to do with their thinking and proceeded to explain how their choices were made. Students seemed to report what made sense to them and what was consistent with their own thinking.

Any effect on performance of the leading criteria probe and principle probe elicitations would not necessarily appear in the item being done. In these two treatments, students first chose their answers and then were asked the questions about whether specific pieces of information affected their choices. So, the elicitation could not have affected their original answer choice. However, students knew they could change their answers at any time, but such changes were made rarely. Also, the elicitation for one item could have affected performance on subsequent items. Students could have predicted on the basis of previous questions that they would be asked whether some specific piece of information in the item affected their choice. Consequently, they might have been more diligent in trying to focus on what was relevant. However, no effects of such a hypothesized increase in diligence were observed. This result is supported by the findings of Phillips (in press) which show that students did no better on a multiple-choice test of inference in reading, which necessarily makes the correct answers available, than they did on a construct-response version of the same test.

Furthermore, in the think aloud and immediate recall elicitations, students knew before they started an item what they would have to do, namely, report all they were thinking in the former treatment and give reasons for their answer choice in the latter. Therefore, these treatments could have affected the original answer choice on the item being done. But no differences between elicitation groups on either performance or quality of thinking were found.

A third reason making the results of this experiment plausible is that effects were sought from a number of directions, but were found in none of them. Among the elicitation groups, there were no differences either in the quality of students' thinking or in the patterns of verbal moves that typified their verbal reports. Between the elicitation and control groups, there were no differences either for performance concurrent with reporting or subsequent to it. It is plausible to think that if differences existed they would have been detected by at least one of these methods.

In addition to the above considerations, an analysis of the statistical power of the experiment, performed using techniques described in Kirk (1968, pp. 9-11, and pp. 107-108), showed <3% chance of a Type II error overall. The analysis requires the calculation of a parameter and the use of charts based upon a procedure by Tang (1938). The parameter is given by:

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

where:

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

n = size of the jth sample

σ_ϵ^2 = error variance.

In the calculation, $(k\text{-}1/n)(MS_{BG} - MS_{WG})$ 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 probability of Type II error was calculated to be <1% for the analyses of verbal reporting and performance and <3% for the analysis of verbal reporting and quality of thinking.

Context-Specific Effects

In the introduction, I limited the study to verbal reports of thinking on multiple-choice tests requiring deliberative thought. Verbal reports of thinking seem useful for validating such tests, because examinees plausibly would have something to say about how they chose their answers. On a test of rote recall or some other automatic process, subjects by definition are unlikely to have access to their thinking. So, collecting verbal reports of thinking does not make sense in this latter context. This intuition is supported by Bereiter and Bird (1985), who also believe that verbal reports of thinking would be most useful in activities requiring deliberative thought. Such activities would include the critical thinking task used in this study and other critical thinking tasks, problem solving and decision making tasks, subject matter tasks requiring deliberation and reflection instead of rote recall, and tests of reading comprehension which require deliberative thinking such as some tests of inference and other higher order processes in reading.

The need and desire to think deliberatively may help explain why different elicitation procedures did not affect thinking in the situation studied in this experiment, but why eyewitness testimony research consistently shows differential effects on the accuracy of verbal reports for different elicitation procedures. Students thought deliberatively on the test because the task required it and, even though the test did not count for school grades, the students wanted to portray themselves as capable people. There is some evidence that subjects in eyewitness testimony experiments may not deliberate about their task in this way. In a critical analysis of eyewitness testimony research, McCloskey and Egeth (1983) contended that while laboratory research suggests that "jurors" place an unwarranted amount of confidence in eyewitness testimony, studies of real jurors do not show this tendency. Real-life jurors tend to be skeptical of evidence and deliberative in their thinking in order to maintain the presumption of innocence. Maintaining a presumption of innocence is not crucial in psychological experiments.

Implications

Verbal reports of thinking would be useful in the validation of multiple-choice tests of deliberative thinking if they could provide evidence for judging whether good thinking was in general associated with choosing keyed answers and poor thinking with choosing unkeyed answers. This study focussed

primarily on one necessary condition for this usefulness to exist, namely, that giving verbal reports does not alter the course of thinking and performance. But even if, as the evidence suggests, they do not alter thinking or performance, they must contain enough information to allow comparisons to be made between the quality of examinees' thinking and their chosen answers.

In fact, the verbal reports of thinking contained a wealth of information useful for rating the quality of subjects' thinking and for diagnosing specific problems with items, such as the presence of misleading expressions, implicit clues, unfamiliar vocabulary, and alternative justifiable answers to the one keyed correct (Norris, in press-a, in press-c). Given the results of this study, it is reasonable to trust this diagnostic information as an accurate representation of problems that would occur with the items taken in paper-and-pencil format.

Multiple-choice tests are popular largely because of their ease of administration and scoring. But the source of this popularity leads to criticisms of them. One criticism is that multiple-choice tests intended to examine deliberative thought and not mere rote recall provide no direct evidence of the reasoning examinees use to choose their answers. On account of this criticism, many educators believe that multiple-choice testing encourages an overemphasis on getting the right answers and undervalues careful reasoning. A systematic procedure for quantifying and using the data in verbal reports of thinking for developing and validating multiple-choice tests can overcome this criticism. Multiple-choice tests could be developed for which the evidence from verbal reports of thinking indicate that, in general, sound thinking is associated with choosing keyed answers and unsound thinking with choosing unkeyed answers (Norris, in press-a). Verbal reports of thinking thus offer the prospect of developing multiple-choice tests which can serve both the desires for efficiency and cost-effectiveness and educational quality.

References

Afflerbach, P., & Johnston, P. (1984). On the use of verbal reports in reading research. *Journal of Reading Behavior, 41,* 307-322.

Anastasi, A. (1988). *Psychological testing.* New York: Macmillan.

Arter, J. A., & Salmon, J. (1987). *Assessing higher order thinking skills: A consumer's guide.* Portland, OR: Northwest Regional Educational Laboratory.

Bereiter, C., & Bird, M. (1985). Use of thinking aloud in identification and teaching of reading comprehension strategies. *Cognition and Instruction, 2,* 131-156.

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,* 269-275.

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

Embretson, S., Schneider, L. M., & Roth, D. L. (1986). Multiple processing strategies and the construct validity of verbal reasoning tests. *Journal of Educational Measurement, 23,* 13-32.

Ennis, R. H., & Norris, S. P. (in press). Critical thinking testing and other critical thinking evaluation: Status, issues, needs. In J. Algina & S. M. Legg (Eds.), *Cognitive assessment of language and mathematics outcomes.* Norwood, NJ: 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 & 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.

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

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*, 585-589.

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.

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

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.

McPeck, J. (1981). *Critical thinking and education*. New York: St. Martin's Press.

Messick, S. (in press). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.). New York: Macmillan.

Norris, S. P. (in press-a). *Verbal reports of thinking and multiple-choice critical thinking test design*. University of Illinois at Champaign-Urbana, Center for the Study of Reading..

Norris, S. P. (in press-b). Informal reasoning assessment: Using verbal reports of thinking to improve multiple-choice test validity. In D. N. Perkins, J. Segal, & J. F. Voss (Eds.), *Informal reasoning and education*. Hillsdale, NJ: Erlbaum.

Norris, S. P. (in press-c). Controlling for background beliefs when developing multiple-choice critical thinking tests. *Educational Measurement*.

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

Norris, S. P., & King, R. (1984). *The design of a critical thinking test on appraising observations*. St. John's, Newfoundland: Memorial University of Newfoundland, Institute for Educational Research and Development. (ERIC Document Reproduction Service No. ED 260 083)

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.

Petrie, H. (1986). Testing for critical thinking. In D. Nyberg (Ed.), *Philosophy of education 1985*. Normal, IL: Philosophy of Education Society.

Phillips, L. M. (in press). *Developing and validating assessments of inference ability in reading comprehension*. Urbana: University of Illinois, Center for the Study of Reading.

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

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

Footnote

[1]I have subsequently concluded that the 3-point thinking score scale was not suitable. To get 3 points, students had to generalize beyond the specific situation of the item by referring to a general principle of critical thinking under which the specific case fell. Hardly any students did this and I now believe that it is pedantic to expect it. Therefore, the effective thinking score range is 0-2 per item, or 0-30 for the 15 items for which students gave verbal reports of their thinking. Thus, students averaged 8.7 on the 30-point scale.

Author Notes

I thank Robert Barcikowski, Robert Crocker, Susan Embretson, and Linda Phillips for helpful comments. Requests for reprints should be directed to Stephen P. Norris, Institute for Educational Research and Development, Memorial University of Newfoundland, St. John's, Newfoundland, Canada, A1B 3X8, or (709) 737-8693.

Table 1

Description of Elicitation Levels

Elicitation Level	Description
Think Aloud Elicitation	Subjects were instructed to report all they were thinking as they worked through an item and to mark their answer on a standardized answer sheet.
Immediate Recall Elicitation	Subjects were asked to mark their answer to an item on a standardized answer sheet and to tell why they chose the answer they did.
Criteria Probe Elicitation	Subjects were asked to mark their answer on a standardized answer sheet and then to tell whether a piece of information pointed out in the item at that time had made any difference to the answer they chose.
Principle Probe Elicitation	Subjects were treated as in the criteria probe group with an additional question asking whether their choice of answer was based upon particular general principles.
No Elicitation (Control)	Subjects were not interviewed, but were instructed to work alone on the test and to mark their answers on a standardized answer sheet.

Table 2

Verbal Report Elicitation Procedures for Item 1

Elicitation	Instructions to Subjects
Think Aloud	Try to tell me all that comes to your mind as you think about this question.
Immediate Recall	Tell me which answer you choose and why you choose that answer.
Criteria Probe	Which answer do you choose? .... Did the fact that Pierre is the map reader affect your choice?
Principle Probe	Which answer do you choose? .... Did the fact that Pierre is the map reader affect your choice? .... If "No," go on to the next item. If "Yes," ask: What difference did it make to your thinking that he is the map reader?

Table 3

Mean Concurrent and Subsequent Performance for Elicitation Level, Interviewer, Sex, and Grade Level

Factor	Level	Mean Concurrent Performance	Mean Subsequent Performance
Elicitation	No Elicitation (Control)	7.8	8.4
	Think Aloud	8.0	8.4
	Immediate Recall	8.3	8.3
	Criteria Probe	7.9	8.6
	Principle Probe	7.6	8.1
Interviewer	A	7.6	8.2
	B	8.2	8.5
Sex	M	7.7	8.0
	F	8.0	8.7
Grade Level	10	7.8	7.8
	11	7.7	8.6
	12	8.1	8.8

Table 4

Mean Thinking Scores for Elicitation Group, Interviewer, Sex, and Grade Level

Factor	Level	Mean Thinking Score
Elicitation 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	M	9.2
	F	8.3
Grade Level	10	8.2
	11	8.6
	12	9.5

Table 5

Frequency of Verbal Moves by Elicitation Group

Verbal Moves	Elicitation Group			
	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
Controlling Attention	26	25	15	19
Interacting with Experimenter	19	9	12	13
Pausing	499	387	424	380