ILLINOIS

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

PRODUCTION NOTE

University of Illinois at
Urbana-Champaign Library
Large-scale Digitization Project, 2007.

Technical Report No. 429

CONTROLLING FOR BACKGROUND BELIEFS WHEN DEVELOPING MULTIPLE-CHOICE CRITICAL THINKING TESTS

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

July 1988

Center for the Study of Reading

TECHNICAL REPORTS

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN
174 Children's Research Center
51 Gerty Drive
Champaign, Illinois 61820

CENTER FOR THE STUDY OF READING

A READING RESEARCH AND EDUCATION CENTER REPORT

Technical Report No. 429

CONTROLLING FOR BACKGROUND BELIEFS WHEN DEVELOPING MULTIPLE-CHOICE CRITICAL THINKING TESTS

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

July 1988

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

The research upon which this publication was based was supported in part by grants from the Social Sciences and Humanities Research Council of Canada, Grant Nos. 418-81-0781 and 410-83-0697. It was completed while the author was a visiting scholar at the Center for the Study of Reading. This paper is forthcoming in Educational Measurement.

This page is intentionally blank.

Abstract

This paper describes briefly a methodology for developing multiple-choice critical thinking tests which attempts to overcome certain problems of validity and fairness facing such tests. The concerns arise for two reasons: (a) it is plausible that for many multiple-choice critical thinking tests it is not differences in critical thinking ability but differences in other factors, such as examinees' background beliefs, that accounts for most variance in test performance; and (b) there is no direct evidence to counter this plausible hypothesis. The proposal is that such direct evidence be gathered during test development by eliciting from samples of students verbal reports of their thinking as they work through trial items. Items would be retained, modified, or discarded according to whether or not critical and uncritical thinking is related, respectively, to choosing keyed and unkeyed answers to the items.

CONTROLLING FOR BACKGROUND BELIEFS WHEN DEVELOPING MULTIPLE-CHOICE CRITICAL THINKING TESTS

During the last decade there has been an increasing interest in teaching critical thinking (Follman, 1987; Resnick, 1987). The interest is motivated, in part, by data showing that school children learn large amounts of information, but learn less well how to analyze, synthesize, and evaluate that information for their own use (National Assessment of Educational Progress, 1985; National Commission on Excellence in Education, 1983).

Concomitant with this growing interest in critical thinking instruction is an increasing desire to test for critical thinking. In meeting this desire there is a heavy reliance on multiple-choice tests. However, many people (e.g., McPeck, 1981; Petrie, 1986) claim that there are inherent flaws in multiple-choice tests of critical thinking. One purported flaw is that such tests cannot be used to distinguish variance in scores due to differences in those background beliefs of examinees which are not part of ability to think critically from variance due to differences in critical thinking ability. For many existing multiple-choice critical thinking tests, this criticism is well founded (Ennis & Norris, in press).

If critical thinking assessment is to succeed, the problem of confounding background beliefs and critical thinking ability when interpreting scores must be solved. In this paper I shall illustrate how this problem with multiple-choice critical thinking tests can be lessened by using verbal reports of examinees' thinking to help develop the tests.

In the first section, I show how multiple-choice testing of critical thinking can lead to a dilemma: Adopting such testing can lead to unfair treatment of students, but disqualifying multiple-choice testing of critical thinking can result in less powerful assessments of critical thinking. The second section illustrates how multiple-choice critical thinking tests can lead to invalid and unfair assessment due to differences in examinees' background beliefs. The third section describes a methodology for using verbal reports of examinees' thinking on trial items to help avoid such invalidity and unfairness.

A Dilemma in Multiple-Choice Critical Thinking Testing

Theories of critical thinking, for instance that of Robert Ennis (1981), generally include standards and criteria for guiding thinking. Only thinking in accord with those standards and criteria is taken to be critical thinking. Nevertheless, the standards and criteria are insufficient by themselves. Knowledge and good judgment are also needed. When thinking about a complex problem, each individual draws upon his or her own background beliefs and sense of good judgment, so each person is likely to reach somewhat different solutions. Since the standards and criteria of critical thinking are not always sufficient to define correct solutions, then more than one solution and approach might reflect critical thinking.

The possibility of more than one good solution to a problem and of more than one good approach to reaching a solution creates difficulties for multiple-choice tests of critical thinking. If the background beliefs of some examinees are different from those of the examiner, then it is possible that, even though the examinees follow the standards and criteria of critical thinking, they will be penalized because they choose answers different from those judged good by the examiner. On the other hand, examinees thinking uncritically might be rewarded merely because they choose the same solutions that the examiner reached.

The above possibilities jeopardize the validity and fairness of multiple-choice critical thinking tests. Unfortunately, there is no easy solution to this problem. On the one hand, the goal of critical thinking instruction is generally focused on teaching students *how* rather than *what* to think, because the critical spirit demands that multiple perspectives be accepted (Paul, 1982; Siegel, 1980, 1988). Thus, opting for test items with only one correct answer seems to introduce a validity-reducing factor into critical

thinking testing. On the other hand, multiple-choice items with one correct answer are one of the best tools in certain evaluation situations, for instance, when the aim is to examine knowledge of the large number of principles for judging the credibility of information or to assess many students. Is there a way out of this dilemma?

I believe there is a way to effect a compromise at the test development stage through the use of verbal reports of students' thinking on trial test items. Before describing that methodology, I shall clarify how differences in background beliefs may lead to differences in performance on multiple-choice critical thinking tests.

Differences in Background Beliefs as Explanations of Performance on Critical Thinking Tests

The effect of differences in background beliefs will be illustrated using items from two commercially available critical thinking tests. The discussion applies to *any* multiple-choice critical thinking tests (with the possible exception of deduction tests), however, since they all are subject to similar sorts of effects.

Briefly, the argument to be made is that for many multiple-choice critical thinking tests variance in performance may be due more to differences in background beliefs than to differences in critical thinking ability, and that at present there is little evidence to indicate the extent to which this may occur. Consequently, there are several possible beneficial effects from taking this issue seriously enough to reexamine existing multiple-choice critical thinking tests: (a) it could lead to empirical results which either support the criticism or exonerate the criticized tests; (b) it could remove some of the suspicion which diminishes people's confidence in such tests and thus jeopardizes their usefulness; and (c) it can force us to alter our test development methodologies in ways that might save multiple-choice tests for use in situations where they are eminently suitable.

The Watson-Glaser Critical Thinking Appraisal

The Watson-Glaser Critical Thinking Appraisal (Watson & Glaser, 1980a), first developed in the late 1930s, is one of the oldest and most widely used critical thinking tests. It has often served as a bench mark for judging the validity of other critical thinking tests and for evaluating the effectiveness of attempts to teach critical thinking. But the documentation available for the test (Watson & Glaser, 1980b) gives no direct evidence that variance in performance on the test is due primarily to differences in critical thinking ability and not to other factors, such as differences in background beliefs which are not part of critical thinking ability. Moreover, a plausible case can be made that several items test for differences in background beliefs, not critical thinking.

Consider Item 6 as an example. For the item, examinees are to read a short passage. They are then given a statement and, on the basis of what they read in the passage, are to judge the statement either True, Probably True, False, or Probably False, or to judge that there is Insufficient Data to make a choice on the truth or falsity of the statement. Here is the passage:

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

Here is the statement to be judged:

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

The answer keyed correct is "Probably True" which, according to the test instructions, means that it is more likely true than false that on some nights it was to Mr. Brown's advantage to keep his pool hall open after 1 a.m. But why is "Probably True" the keyed answer? There is no rationale provided in the test manual. In addition, the manual provides no direct evidence that examinees choosing the keyed answer generally think critically and that those choosing an unkeyed answer generally think uncritically. But this is the relationship which *must* exist if the item is to differentiate among examinees on the basis of their critical thinking ability.

Given the lack of direct evidence, are there logical reasons for believing that the item works the way it should? That is, is it plausible that generally when examinees choose the keyed answer they do so because they have thought critically and that generally when they choose an unkeyed answer they have thought uncritically? Plausibility in this case, I submit, is inversely proportional to the number of plausible ways that have not been eliminated by evidence for examinees to think well and choose unkeyed answers and to think poorly and choose the keyed response. The following argument, based upon one by Norris and Ennis (in press), shows that the Watson-Glaser test cannot meet this standard of plausibility.

Suppose an examinee recognized the possibility that Brown was not telling the truth when admitting guilt. Maybe it was Brown's son who kept the pool hall open and, although he disagreed with his son's action, Brown preferred to take the blame himself rather than see his son face the charges. On the other hand, an examinee might think that Brown was a victim in a cover-up and that admitting guilt to an offense he did not commit was a way of channelling money to crooked municipal government officials. Or, an examinee might think that Brown was telling the truth but was suffering from a severe shock which provoked him to do things that were not to his advantage. Another examinee might consider that Brown kept his pool hall open late to protest what he considered an unfair ordinance which allowed only some establishments to remain open after 1 a.m. He did not think it was to his advantage to protest, but did so on principle.

If an examinee's background beliefs led him or her to assume any of the above possibilities, then the examinee would be justified in choosing "Probably False" as the correct answer. If an examinee thought of a number of these possibilities, but could not decide among them on the basis of the information given, then that examinee would be justified in choosing "Insufficient Data" as the correct answer. In both cases, the examinees would be marked wrong even though they thought well. But because of the multiple-choice format we would not have known how well they thought.

There is no available evidence on which possibilities actually do come to the minds of examinees for whom the Watson-Glaser test is designed (junior high school through college level). As a result, examinees' choices of answers do not provide sufficient information for deciding whether or not they are thinking critically. Therefore, if we are to use the test as designed, we must rely by default when rating examinees' critical thinking on whatever reasoning led the test developers to choose "Probably True" as the keyed response. We are not told what this reasoning is and there is no evidence of the extent to which the reasoning of examinees who choose the keyed response matches that of the test developers.

We know for sure that thinking critically *can* lead justifiably to different answer choices depending upon the background beliefs used. But since there is no evidence on how examinees tend to think when they reason through individual items on the test, and since there is no information on the reasoning which supports the keyed responses to those items, there is no reason to believe that *in general* when examinees choose keyed responses they think critically and when they choose unkeyed responses they think uncritically.

This criticism cannot be countered by arguing that background belief effects will wash out in the averages over all items on the test, because many items on the test are subject to the same sort of criticism. Nor can the criticism be countered by pointing to the large amount of correlation data

relating performance on the Watson-Glaser test to other variables (Watson & Glaser, 1980b). This data provides evidence on the convergent and discriminant validity of the test, but it is not compelling when the same data used to elaborate the nomothetic span of a test is also used to clarify the construct the test measures in the first place (Embretson, Schneider, & Roth, 1986). This is especially true for the construct of critical thinking. Conjectures about how critical thinking *should* correlate with other variables are quite untrustworthy, given the status of the theory of the construct. Therefore, inferring whether or not a test measures critical thinking from how it correlates with other variables is doubly risky.

Therefore, we do not know the extent to which examinees are unfairly penalized for choosing unkeyed responses, even though they have thought critically, or the extent to which examinees are unfairly rewarded for merely choosing keyed responses while using no critical thought at all. That is, the prevalence of the problem is not known, because there has been no systematic investigation of it.

Test on Appraising Observations

The Test on Appraising Observations (Norris & King, 1983) focuses on one aspect of critical thinking, the ability to evaluate statements of observation. This focus on a single aspect distinguishes it from the Watson-Glaser test, which examines several aspects of critical thinking. But the Test on Appraising Observations is a multiple-choice test, so it is subject to the same sort of potential problems from differences in examinees' background beliefs. Therefore, the development methodology of the Test on Appraising Observations, which was designed to minimize these problems, is relevant to the development of the Watson-Glaser and other multiple-choice critical thinking tests.

The purpose of the test is to assess knowledge of various principles for judging the credibility of reports of observations. In Part A, items are cast in the context of a traffic accident. Witnesses and people who were involved in the accident report what they observed happening. In each item, two underlined reports are presented and the task is to decide which, if either, of the reports is more believable.

Consider Item 9. In it, Ms. Vernon and Martine, both witnesses to the accident and drivers of nearby but uninvolved cars, report on cars they saw going through a stop sign.

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

Examinees are told to choose between the underlined statements. The answer keyed correct is that Vernon's observation is more believable, because being a fancy blue sports car is taken to be more salient than having twin headlights. The item is intended to test knowledge of the Principle of Observational Salience: Observations of more salient features of events tend to be more credible than observations of less salient features. Features of events are salient to the degree that they are extraodinary, colorful, interesting, and novel, and not salient to the degree that they are routine and commonplace (Loftus, 1979; Nisbett & Ross, 1980).

But do examinees' choices of answers indicate their knowledge of the Principle of Observational Salience? Consider an examinee who knows the critical thinking principle, but believes that having twin headlights is a more salient feature than being a fancy blue sports car. Based on this belief, the examinee would be justified in choosing Martine's statement as more believable. Or, suppose an examinee knows the principle, but believes that neither feature is more salient. That examinee is justified in deciding that neither statement is more believable. Imagine two other examinees who know the principle and believe that being a fancy blue sports car is more salient in the daytime, but that having twin headlights is a more salient feature at night. If one examinee imagines the situation to be at night (justifiably choosing Martine's statement as more believable) and the other imagines it to be

day (justifiably choosing Vernon's statement), then one examinee will choose the keyed response and the other an unkeyed response, even though both know the principle being tested.

The intent of the item is to differentiate between those who know and do not know the Principle of Observational Salience, not to differentiate between those who know and do not know whether being a fancy blue sports car or having twin headlights is more salient, or between those who imagine it is night and those who imagine it is day. However, on a multiple-choice item where only choice of answer is revealed, it is difficult to know on which basis differentiation among examinees is really being made. Thus, the test faces the same potential difficulty as the Watson-Glaser and all other multiple-choice critical thinking tests.

However, the methodology used to develop the Test on Appraising Observations, which is described in the following section, provides evidence that greater than 95% of the high school students for whom the test is designed assume that the situation takes place during daytime. In addition, the methodology provides evidence that fewer than 10% of high school students who know the principle being tested assume that having twin headlights is a more salient feature than being a fancy blue sports car. Furthermore, the evidence shows that there is a correlation of .87 between thinking critically on the item and choosing the keyed response.

Similar evidence is available for each item on the test, although of course the numbers are not exactly the same. Thus, in contrast to the Watson-Glaser and most other critical thinking tests, there is *direct* evidence that the test differentiates primarily on the basis of differences in critical thinking and not some other factors, such as differences in background beliefs.

A Methodology for Developing Multiple-Choice Critical Thinking Tests

Trial versions of the Test on Appraising Observations were vetted by asking samples of students to think aloud as they worked on the items (Norris & King, 1984). Items were retained, modified, or discarded according to whether or not it was critical thinking and not some other factors, such as background beliefs, which was the major contributor to differences in scores. This procedure was repeated with revised test versions until the average correlation between thinking critically and choosing the keyed response was greater than .70 across all 50 items.[1]

High school students took the trial versions in a one-on-one, tape-recorded interview format with one of the test developers. The interviews were conducted so as to be as non-leading as possible. The aim was to try to elicit from students reasoning that was not different in substantive ways from the reasoning they would have done had they taken the test in the normal paper-and-pencil format. The interview approach has been shown subsequently not to change substantively the course of examinees' thinking (Norris, in press).

First, the directions of the test were made clear to students. They were then asked to read the first item aloud, to mark their answer-choice on an answer sheet, and to say all that they were thinking as they chose their answer. At this stage of the interview, the interviewer interrupted students only to probe for ambiguous references and to check for reading errors. If students asked for additional information, they were told that no information other than that in the test could be given. When students had finished talking about the item, the interviewer had the option to pose questions before the students were asked to proceed to the next item. These questions were more leading and requested the specific reasons for students' choices of answers when these reasons were not made clear in what they had said. The procedure was repeated for subsequent items.

For a given trial version of the test, about 50 students were interviewed. Each student was asked to think aloud on about one-fourth of the items and to do the remaining items in a paper-and-pencil sitting. Thus, for any item, from 12 to 15 verbal reports of thinking were obtained. These reports were

transcribed and then analyzed for the quality of the critical thinking they portrayed. The analysis involved studying each student's report and assigning a Thinking Score from 0 to 3 for each item to indicate quality of the student's thinking. The thinking score for each item was based on the degree to which the student's thinking matched an ideal model of thinking on that item. Thinking scores were assigned independently of the answer chosen, so it was possible for a student to obtain a high thinking score on an item, but choose an unkeyed answer, or to obtain a low thinking score but choose the keyed answer.

Thus, for each item on the test there were two sets of scores. The set of thinking scores (0s, 1s, 2s, or 3s) represented the quality of each student's verbal report of thinking on that item. The set of performance scores (0s or 1s) represented whether or not the students had chosen the keyed answer. These two sets of scores were correlated. A high correlation represents a high correspondence between thinking critically or uncritically on a given item and choosing, respectively, the keyed response or unkeyed response. Thus, the correlations provide direct evidence of how well items are working.

When correlations were low, items were revised and retried using the same format. The verbal reports were very useful in making these revisions, because they often made quite clear why an item was not working as desired: There might have been an ambiguity in wording; students might not have understood what a particular expression meant; or students might have used background beliefs not part of their critical thinking ability which were different from those used by the test developers in choosing the keyed response.

An example of changes made to the directions will illustrate how the methodology worked. Recall that items on the test are cast in the context of a traffic accident. In one version of the test, the directions contained the names of all characters and what they were doing when the accident occurred. The rationale was that the list of characters and roles would help examinees keep the information straight. But for that test version the correlation between thinking and performance scores for the first several items was too low. The verbal reports showed that many examinees used the information in the directions to answer these questions. To illustrate, in one item, two characters gave conflicting reports about how many cars were at the intersection. One character was more alert and therefore a more credible witness. Coincidentally, that character reported that there were three cars *at the intersection*, the same number that the directions said were *involved in the accident*. The verbal reports showed that several examinees did not consider the alertness of the witnesses or any other relevant feature of the situation, but simply cited the number of cars mentioned in the directions as their answer. These students equated uncritically the number of cars at the intersection with the number involved in the accident, but nevertheless chose the keyed response. This problem with the test was made quite prominent by the verbal reporting methodology.

Discussion and Concluding Remarks

I have given only a brief sketch of how the use of verbal reports of thinking on trial test versions can help provide evidence on the validity of multiple-choice critical thinking tests. The relevance of verbal reports of thinking to test construction has been suggested by several testing specialists (e.g., Anastasi, 1988; Cronbach, 1971; Haney & Scott, 1987; Messick, in press), but the technique has been used rarely (Norris, in press). However, verbal reports of thinking are *particularly* relevant to the development of multiple-choice critical thinking tests, because satisfying the purposes of such tests demands direct evidence on the thinking processes students follow when taking them. In tests designed to distinguish students who know certain pieces of factual information from those who do not, then knowing the thinking processes of examinees might not be crucial. But when thinking processes are the focus of the evaluation, inferences about students' abilities based merely on the answers they choose tend to be untrustworthy.

The verbal report methodology does have some shortcomings. First, there is the problem of generalizing to examinees other than those whose verbal reports were obtained. Generalizability is

always a problem in research, but we do not know the extent of the problem for critical thinking testing. In particular, there is no good evidence on the extent to which different subgroups bring different background beliefs to bear on the same problems. Thus, it is still not known how much of a solution the proposed methodology effects. For example, only high school students were involved in the verbal report studies of the Test on Appraising Observations. So we do not know how valid the test is for junior high school students, college students, or high school students from different places.

What are the alternatives to the standardized multiple-choice tests currently available? One alternative is to avoid multiple-choice testing altogether, maybe by using constructed-response tests which require essays or short answers. In their essays and short answers, examinees might reveal the background beliefs upon which their thinking is based. Examiners could then take this information into account in making judgments about examinees' levels of critical thinking. However, constructed-response testing does not provide complete assurance. There is no window into examinees' brains which shows all they are thinking or all of the basis for their decisions. Examinees likely do not know all of these things themselves (Nisbett & Wilson, 1977). In addition, constructed-response testing raises other concerns, such as low interrater reliability and the inability to adequately cover a wide range of critical thinking abilities and dispositions in a reasonable time (Norris, 1986). If, for example, evaluation of examinees' ability to appraise observations is the concern, then it is difficult to imagine how knowledge of all the principles of observation appraisal could be assessed in a constructed-response test of reasonable length.

Another alternative is to mix standard multiple-choice formats with other formats. For example, multiple-choice items could be supplemented by asking examinees to provide reasons for the answers they choose. So as not to turn a multiple-choice test entirely into a constructed-response test, reasons might be sought for only some of the items. Such an approach would provide an indication of the background beliefs examinees were using in their thinking and the examiner could take these beliefs into account in assessing their critical thinking.

A third alternative might be to base critical thinking tests only on school subject matter which students are supposed to have studied, instead of on general knowledge as found in most current tests. Doubtless, not all students will have learned the particular body of knowledge to the same degree, so differences in background beliefs will continue to cause variance in scores. However, the influence of differences in background beliefs may be lessened and, even if not, it would not pose the same issues of fairness that arise when using critical thinking tests based upon general knowledge. If students perform poorly on a critical thinking test in science because they have not learned the required science content, then, barring poor instruction or other extenuating circumstances, it is not unfair to mark them down. However, there remains the question whether critical thinking or science content was being tested. This validity issue would still need to be addressed.

In addition, critical thinking testing based on school subject matter may not be the best way to determine whether critical thinking has generalized to problems outside of school subjects. Much of the justification for teaching critical thinking is based on an expected generalizability to everyday life situations outside the school and school subject matter, so it is important to test for how much this expectation is realized.

Minimizing problems arising from differences in background beliefs when testing for critical thinking should be an important concern for those involved in critical thinking evaluation. First, there is the concern for validity. If scores on tests are to be interpreted as measures of critical thinking ability, then it is necessary to reduce as much as possible the effects of differences in background beliefs. Second, there is a concern for fairness and, as Messick (1975) argued over a decade ago, validity and fairness go hand in hand. If we believe our critical thinking tests are valid, but we are really differentiating among students on the basis of background beliefs unrelated to critical thinking, then there is a risk of treating them unfairly. There is a risk of unfairly penalizing students who are thinking critically but who do not

have the appropriate background beliefs, and a risk of unfairly rewarding students who are not thinking critically but who, nevertheless, have those background beliefs which enable them to perform correctly.

The desire to teach critical thinking places many new demands on educators. One new demand is that test development practices will need alteration in order to allow for the diversity of opinion and approaches to problems which critical thinking encourages. The alterations may include the adoption of time consuming development methodologies such as the one described here. But the ideal of critical thinking is worth the effort. Otherwise, we are left with the worn-out and educationally indefensible emphasis on memorization of factual information, rote recall, and pat answers.

References

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

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

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. (1981). Rational thinking and educational practice. In J. F. Soltis (Ed.), *Philosophy of education,* 80th yearbook of the National Society for the Study of Education, Vol. 1. Chicago: The National Society for the Study of Education.

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

Follman, J. (1987). Contemporary critical thinking activity update. *CT News, 6*(1), 4-7.

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

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

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

Messick, S. (1975). The standard problem. *The American Psychologist, 30,* 955-966.

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

National Assessment of Educational Progress (1985). *The reading report card, progress toward excellence in our schools, trends in reading over four national assessments.* Princeton, NJ: Educational Testing Service.

National Commission on Excellence in Education (1983). *A nation at risk.* Washington, DC: Superintendent of Documents.

Nisbett, R., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment.* Englewood Cliffs, NJ: Prentice-Hall.

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

Norris, S. P. (1986). Evaluating critical thinking ability. *The History and Social Science Teacher, 21,* 135-146.

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

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

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

Paul, R. (1982). Teaching critical thinking in the strong sense: A focus on self-deception, world views, and a dialectical model of analysis. *Informal Logic Newsletter, 4*(2), 2-7.

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

Resnick, L. B. (1987). *Education and learning to think*. Washington, DC: National Academy Press.

Siegel, H. (1980). Critical thinking as an educational ideal. *Educational Forum, 45*(1), 7-23.

Siegel, H. (1988). *Educating reason: Rationality, critical thinking, and education*. London: Routledge.

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

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

Author Notes

I thank Linda Phillips and Walter Haney for helpful comments.

Requests for reprints should be sent to Stephen P. Norris, Institute for Educational Research and Development, Memorial University of Newfoundland, St. John's, Newfoundland, Canada, A1B 3X8, (709) 737-8693.

Footnote

[1]Correlation coefficients were not actually used. The biserial correlation coefficient was the most suitable estimate of correlation given the nature of the data, but it is subject to distortion from a variety of factors. Consequently, some correlations were greater than 1.0 and, hence, not interpretable. A statistic, called a Thinking/Performance Index, was developed and used in place of the correlations. The T/P Index is a measure of the net positive evidence available for an item from the interview data. It is described more fully in Norris and King (1984).