Can We Test Validly for Critical Thinking?

STEPHEN P. NORRIS

This paper focuses first on two questions: (a) Is critical thinking generalizable? and (b) What is a critical thinking disposition? It is argued that the controversial nature of these questions limits the ability to judge the validity of critical thinking testing. Following this discussion, the paper outlines a series of procedures that can help increase the validity of one fundamental aspect of critical thinking testing—multiple-choice testing of credibility judgment. The procedures rely on verbal reports of examinees' thinking on items to gain direct evidence on the reasons for their answer choices. It is recognized that multiple-choice tests cannot test all important aspects of critical thinking, but that improving multiple-choice tests where they are applicable can have important practical and scientific implications.

Educational Researcher, Vol. 18, No. 9, pp. 21-26

This paper focuses first on two questions: (a) Is critical thinking generalizable? and (b) What is a critical thinking disposition? It is argued that the controversial nature of these questions limits the ability to judge the validity of critical thinking testing. Following this discussion, the paper outlines a series of procedures that can help increase the validity of one fundamental aspect of critical thinking testing—multiple-choice testing of credibility judgment. The procedures rely on verbal reports of examinees' thinking on items to gain direct evidence on the reasons for their answer choices. It is recognized that multiple-choice tests cannot test all important aspects of critical thinking, but that improving multiple-choice tests where they are applicable can have important practical and scientific implications.

Typical characterizations of critical thinking as an educational goal lead to problems in testing for critical thinking. According to one definition, critical thinking is "reasonable and reflective thinking that is focused upon deciding what to believe or do" (Norris & Ennis, 1989, p. 1). The breadth of this definition makes it difficult to decide on the coverage of a critical thinking test, since decisions about belief and action are all the decisions that we make. Furthermore, it is difficult to identify objective standards for judging the reasonableness and reflectiveness of examinees' thinking. According to another definition, critical thinkers are "appropriately moved by reasons" (Siegel, 1988, p. 32), that is, are able to think critically and are disposed to do so. Since human disposition involves the will, a terribly perplexing notion, it is not clear how to devise a test of the disposition to think critically.

This paper focuses on such problems facing critical thinking testing. In the first two parts, I examine two questions central to understanding critical thinking: (a) Is critical thinking generalizable? and (b) What is a disposition to think critically? I explain how the controversy surrounding these questions inhibits our ability to judge the validity of critical thinking testing. In the third part, I discuss some examinee characteristics that influence the validity of critical thinking testing and focus on an approach for reducing these influences in tests of credibility judgment. I chose to examine tests of credibility judgment on the grounds that this aspect of critical thinking is one of the most important and best understood. The goal is to produce improved tests that can serve as baseline instruments in a bootstrapping effort that can lead to partial resolutions of the above two questions, thence to even more improved tests, thence to better resolution of the questions, and so on.

The Generalizability of Critical Thinking

The question of the generalizability of critical thinking abilities and dispositions has been the subject of much debate (Ennis, 1989; Glaser, 1984; McPeck, 1981; Norris, 1985a, 1988a; Orton, 1988; Paul, 1985; Perkins & Salomon, 1989; Resnick, 1987). The issue is confusing, because the expression "the generalizability of critical thinking" has at least two senses—an epistemological sense and a psychological sense.

Epistemological Generalizability

According to epistemological generalizability, there are principles and standards of critical thinking that are applicable to many subjects. A corollary is that in different subjects the same "sorts of things count as good reasons" (Ennis, 1989, p. 7). A proponent of epistemological generalizability might hold that, for example, the following principle applies to all cases of inductive explanation: "An explanatory hypothesis is justified to the extent that . . . it is not inconsistent with any evidence" (Ennis, 1981, p. 174). As another example, the person might hold that a conflict of interest reduces one's credibility as a source of information, no matter what the subject matter.

The strongest critics of epistemological generalizability maintain that "each field of inquiry has its own particular epistemology" (McPeck, 1981, p. 33)—the standards for determining knowledge differ from field to field. Thus, the standards demanding that explanations be consistent with the evidence and that sources of information not be in a conflict of interest may not apply, or may even be transgressed, in certain fields.

An intuitive case can be made for each side of the issue. In favor of epistemological generalizability, the principle of inductive reasoning cited above seems widely applicable. It seems applicable to explanations of evidence gathered to determine the guilt of a defendant in a court of law, the effectiveness of innovative educational practices, the need for light to have a medium of transmission, the relationship between interest rates and unemployment, and the interpretation of a line of poetry.

On the other side, it seems plausible that the standards for determining knowledge might differ from mathematics, to social science, to archaeology, to law. In mathematics, deductive proof is the usual standard of acceptance. In social science, statistical probabilities are the norm. Archaeology seeks historical narratives that explain and bring coherence to evidence, often fortuitously discovered, found at scattered places throughout the earth. The law relies on precedent and case-by-case reasoning, wherein analogous cases are used to help decide the current one.

Examining intuitively plausible cases
on each side is, of course, only part of the debate over epistemological generalizability. My purpose, however, is not to resolve the debate or to explain it in all its details, but to show how, until the debate is resolved, testing validly for critical thinking will be problematic. Examining just about any critical thinking test will illustrate the problem. Let us consider the Watson-Glaser Critical Thinking Appraisal (Watson & Glaser, 1980a). The fifth section of the test, "Evaluation of Arguments," requires examinees to distinguish "between arguments that are strong and relevant and those that are weak and irrelevant to a particular question at issue" (Watson & Glaser, 1980b, p. 2). Three issues are presented in the section: (a) whether a strong labor party would promote the welfare of the people of the United States; (b) whether the United States' defense department should inform the public of its anticipated scientific research; and (c) whether students should be able to receive religious instruction during school hours.

What would it mean for a student to do well on this section of the Watson-Glaser test? If epistemological generalizability is correct, good performance would be evidence that the student knows the critical thinking principles and standards for distinguishing strong and relevant arguments from weak and irrelevant ones, whatever the content of the arguments. Note that the student's good performance would not sanction the inference that the student has all the ability necessary to distinguish strong and relevant arguments from weak and irrelevant ones whatever the content. The good performance only sanctions the inference that the student has the critical thinking ability necessary for the task, because it is consistent with epistemological generalizability that subject-specific knowledge outside critical thinking is necessary to evaluate arguments in different subjects and that, without such extra-critical-thinking knowledge, critical thinking principles and standards are not sufficient (Ennis, 1989). If epistemological generalizability is incorrect, then the inference that the student has the critical thinking abilities needed to evaluate arguments in subjects other than those examined on the test would not be sanctioned, because the anti-epistemological-generalizability view is that the standards of argument evaluation differ from subject to subject. Therefore, the unsettled debate over the correctness of epistemological generalizability makes it difficult to test validly for critical thinking because we often do not know what inferences from test scores to sanction.

Psychological Generalizability

Psychological generalizability is the view that people actually apply critical thinking learned in one subject to thinking in another. Maintaining that such transfer occurs entails epistemological generalizability, because there is nothing to transfer if there are no principles and standards of critical thinking that apply generally. Claims about the application of critical thinking learned in one subject to thinking in another subject also presuppose a distinction between critical thinking that transcends particular subjects and subject-specific knowledge that is necessary to think critically in particular subjects. There is, however, no satisfactory account of how to draw this distinction in practice (Ennis, 1989), so in particular cases it is often difficult to know how to interpret data. Did transfer of critical thinking occur or did individuals only recall their subject-specific knowledge? Did transfer of critical thinking fail to occur or did individuals lack the subject-specific knowledge necessary for transfer?

These unresolved issues of psychological generalizability interact with the questions over epistemological generalizability. Assuming epistemological generalizability is correct, lack of knowledge about psychological generalizability still would hamper interpretations of scores on critical thinking tests, because we would not know the meaning of good performance on a test (which necessarily must use some specific subject matter) for critical thinking performance in other subjects. If epistemological generalizability is false, then transfer is logically impossible, so the issues of psychological generalizability would disappear.

The Disposition to Think Critically

People can have abilities they do not use (Baron, 1985; Bereiter & Scardamalia, 1987; Perkins, 1985). However, simply possessing critical thinking abilities is not an adequate educational attainment—the abilities must be used appropriately. Therefore, normative theorists of critical thinking claim that to be a critical thinker one must have both abilities and dispositions (Ennis, 1981; Norris, 1985b; Siegel, 1988). Critical thinkers are disposed to seek reasons, try to be well informed, use credible sources and mention them, look for alternatives, consider seriously points of view other than their own, withhold judgment when the evidence and reasons are insufficient, seek as much precision as the subject permits, among other activities (Norris & Ennis, 1989).

Furthermore, in order for critical thinking to be an educational ideal, critical thinking dispositions must be more than tendencies to behave—they must be tendencies to behave that are justified and motivated by educational norms (Siegel, 1988). Some students may tend to seek alternatives only because it pleases their teachers, or only because it leads to high grades. Others may raise alternatives only because they love to quarrel. The beliefs, values, and desires underlying these students' tendencies cannot justify or motivate their tendencies in an educationally defensible way. That it pleases the teacher, leads to high grades, or starts a quarrel are not educationally sound reasons for seeking alternatives. Therefore, the students do not have the critical thinking disposition to seek alternatives.

Siegel (1988) has offered several educationally defensible reasons for thinking critically. First, thinking critically entails respect for other persons, because to think critically one must consider seriously the points of view of other people and must respect their right to challenge one's own points of view. Second, thinking critically empowers people to be self-sufficient in choosing and following the way of life that best suits their interests and capacities. Third, critical thinking, conceived as the appropriate use of reasons, enables an understanding of the various disciplines, because to understand the disciplines one must understand the role of reasons in them and how those reasons are assessed. Fourth, critical thinking enables people to partake rationally in the decision making of a democratic society, thus helping to sustain a democratic way of life. According to Siegel, critical thinking that arises from such beliefs, desires, and values as these is justified and motivated in educationally satisfactory ways. His defense of this claim cannot be presented in this paper.
The above analysis of critical thinking dispositions has several implications for testing validly for critical thinking. In particular, there is the question of how to interpret behavior in and of itself. Consider an example: In one section of a constructed-response critical thinking test (Norris & Ryan, 1988), examinees are asked to imagine that they are on the fictitious planet Zed, searching for living creatures. They are told what happens each day, and are instructed to keep in mind the search for living creatures while writing what they are thinking about the things that happened on that and previous days, and what they plan to do on their search because those things happened. Responses are scored positively for, among other things, offering alternative explanations of events that happened and alternative courses of action that might be followed. However, examinees are not told during the test what criteria will be used to grade their responses.

How should we interpret an examinee’s response that offers alternatives? Examinees do not have to give reasons for what they write, so the response does not show whether the examinee understands that seeking alternatives is a route to knowledge. Maybe the examinee sees that writing possibilities is just a way to fill the space.

Suppose, on the other hand, an examinee does not offer alternative explanations of the events that are described. Is this evidence that the examinee lacks the critical thinking disposition to seek alternatives? Maybe the examinee does not have sufficient subject-specific knowledge to think of alternative explanations and courses of action, but knows the value of alternatives in the pursuit of knowledge. Maybe the examinee sees that writing possibilities is a lesson to fill the space.

Current critical thinking tests are unable to deal with such issues. There are no detailed normative theories of the beliefs, desires, and values that ought to underlie critical thinking dispositions, though Siegel has proposed a beginning sketch. In addition, there is insufficient empirical knowledge of the beliefs, desires, and values that actually do underlie thinking, and of the relationship between seemingly general dispositions and the acquisition of knowledge in subject-specific contexts (Perkins & Salomon, 1989). Without such normative and empirical knowledge, it is not clear how to construct better tests of critical thinking dispositions or how to interpret the evidence from existing tests in terms of critical thinking dispositions.

Note that the generalizability issues discussed in the previous section and the disposition issues discussed in this section are related. We can wonder whether critical thinking dispositions, as well as critical thinking abilities, are generalizable. Some of the previous discussion in this section presupposed (for argument’s sake) the epistemological generalizability of dispositions when, for instance, the claim was made that a person may have a disposition to offer alternatives but not do it in a situation because of lack of subject-specific knowledge. Thus, there is another intersection—between dispositions to act and one’s knowledge—that has not been well explored.

### Improving Tests of Credibility Judgment

**Critical thinking theory is a normative theory about how one should go about thinking (Ennis, 1981; Siegel, 1988).** As such, it provides standards and criteria for gaining, assessing, and using information; for making and judging inductive and deductive inferences; and for value judging. As a result of this conception, testing for critical thinking is often seen as an attempt to determine whether examinees are following in their thinking the standards and criteria that have been laid down. When constrained this way, testing validly for critical thinking requires evidence on the process of examinees’ thinking. This requirement has led to criticisms of multiple-choice critical thinking testing on the grounds that it can never give strong evidence on process (McPeck, 1981; Petrie, 1986).

Nevertheless, multiple-choice critical thinking tests are the most popular (Norris & Ennis, 1989) and, because of the economies they afford, are likely to remain popular. Thus, it would be significant if it could be shown that multiple-choice testing of at least some aspects of critical thinking need not be invalid. For the remainder of this paper, therefore, I shall focus on multiple-choice critical thinking testing. Specifically, the focus will be on testing for credibility judgment. Judging the credibility of information is fundamental to thinking critically. In addition, criteria and standards for credibility judgment have been well developed and defended in the field, making this area, next to deduction, the easiest aspect of critical thinking in which to make multiple-choice items. Even, here, however, it is not always possible to assign unconditionally correct answers. Therefore, significant threats remain to the validity of tests that credit one and only one answer as correct.

### Factors Influencing the Validity of Credibility Judgment Testing

Several factors can render invalid multiple-choice credibility judgment testing that accepts one and only one answer as correct, including differences in examinees’ degrees of critical thinking sophistication, extra-critical-thinking empirical beliefs, assumptions made during test taking, and political and religious ideologies (Ennis & Norris, in press; Norris, 1988b, in press-b). A methodology for improving credibility judgment tests would have to provide a means of lessening the effects of these factors.

Consider the effect of the empirical-belief factor on the credibility judgment section of the Cornell Critical Thinking Test Level X (Ennis & Millman, 1985a). The test describes the activities of a team of explorers that has landed on a newly discovered planet, Nicoma, in order to search for a team that had arrived earlier but had not been contacted for 2 years. When exploring the area around their landing site, they find some water and discuss whether it is potable. In the following item, the task is to decide which, if either, of the two italicized statements is more credible.

There are three options: the first italicized statement is more credible, the second is more credible, or they are equally credible.

A. The health officer says, “This water is safe to drink.”
B. Several others are soldiers. One of them says, “This water is not safe.”
C. A and B are equally believable.

The item is designed to test for knowledge of the expertise criterion—that people speaking in their areas of expertise tend to be more credible than non-experts. (The justification of this criterion and others that are tested on the Cornell cannot be given here. The criteria find support in judicial practice,
psychological research on eyewitness testimony, and theories of memory creation and retrieval. Ennis (1981) and Norris and King (1984) provide some of the justification.) The keyed answer is "A," the health officer’s statement is more credible, on the grounds that health officers have more expertise than soldiers on the potability of water. Suppose, however, that an examinee believes that soldiers’ training makes them expert judges of water safety. Such an examinee might choose "C" as the answer, on the grounds that the health officer and soldier are equally expert. The examinee would be marked wrong, but would have used the expertise criterion in choosing an answer. Since the Cornell is intended to test knowledge and application of that critical thinking criterion, and not extra-critical-thinking empirical beliefs about soldiers’ training, problems of validity and fairness arise to the extent that differences in scores are due to differences in such empirical beliefs.

In general, the issue is as follows. A key represents a test maker’s judgment of the correct answers. To decide on the correct answers for a credibility judgment test, the test maker must take into account factors other than criteria for judging credibility, including background empirical beliefs and political and religious ideologies that reasonably could be expected to be held by examinees, and assumptions that examinees would likely make. This means that test makers’ judgments are based on extra-critical-thinking factors that can differ from those on which test takers base their judgments. But examinees should not be penalized on critical thinking tests for taking into account different, but reasonable, extra-critical-thinking factors, nor rewarded merely for taking into account the same factors as the test makers.

Unfortunately, for most commercially available multiple-choice critical thinking tests, there is no evidence on what examinees consider. The aim of the next section of the paper is to outline a procedure for collecting information on the factors that examinees consider and for deciding whether the key unfairly penalizes or rewards too many examinees from given audiences.

Controlling for Extra-Critical-Thinking Empirical Beliefs

One approach for reducing the effects of the above factors would be to use other than strictly multiple-choice tests, possibly tests with multiple-choice items followed by requests for justifications of answers chosen (Ennis & Norris, in press-a). Another approach would continue to use multiple-choice tests, but would study the extra-critical-thinking empirical beliefs (to focus on that factor) that examinees bring to bear on items. Evidence on examinees’ empirical beliefs could be gathered by asking them to write their justifications for answers, by experiments that studied examinees’ performance by controlling for or manipulating their empirical beliefs, or by asking examinees to think aloud while working on items. Ideally, all of these approaches would be used, as each would likely yield somewhat different information on empirical beliefs. Items that discriminated unfairly against a significant proportion of the intended audience, or against an identifiable group that one was especially concerned to treat fairly, could be either eliminated or modified. In the remainder of this section, I shall concentrate only on the approach of asking examinees to think aloud while working on items, for this is an approach I have used with some success.

Verbal reports of thinking. Verbal reports of thinking have found some use in test construction (Haney & Scott, 1987), but rarely in the construction of critical thinking tests. The limited use in critical thinking testing is unfortunate because the approach seems particularly appropriate in that context. There is generally a deeper concern among critical thinking advocates with how examinees justify their answers to problems than with their actual answer choices. Asking examinees to think aloud while working on problems provides one of the most direct sources of evidence on the justifications that examinees use.

In addition to this argument that the approach seems appropriate on the surface, Norris (in press) reports evidence that thinking aloud while working on a multiple-choice critical thinking test did not alter the course of examinees’ thinking or performance compared to paper-and-pencil test taking. Phillips (1989) reports similar results for an inference test in reading comprehension. Such studies provide evidence that the information in verbal reports is relevant to the validity of multiple-choice critical thinking tests used in paper-and-pencil format.

Eliciting verbal reports on credibility items. Credibility judgment tests typically test for knowledge and application of the criteria for making credibility judgments. The item from the Cornell test discussed previously tested the expertise criterion. A complete set of criteria, including ones that refer to conflict of interest, agreement with established sources, and use of established procedures, can be found in Ennis (1981) with explanation and illustration. Each item should test for only one criterion in order to effect the level of control required to interpret scores in terms of examinees’ knowledge of the criteria and their application. Gaining this control means forfeiting the ability to test the important critical thinking goal of using good judgment to weight and balance competing criteria. But it is probably unrealistic to attempt multiple-choice testing of the use of good judgment.

Items should require comparative judgments of statements (Norris & Ennis, 1989) because, without the wealth of information contained in a real-life situation, it is not possible to make absolute judgments of credibility. A comparative judgment is required by the above example item from the Cornell test, because examinees are to choose which, if either, of two statements is more credible.

Having developed a trial set of items meeting the above characteristics, tape-recorded verbal reports of thinking can be collected by asking examinees to think aloud while working on the items, by asking them to justify their answer choices, or by asking them to tell the role in their thinking of specific pieces of information in the items. Norris (in press-a) found that each of these approaches resulted in the same average performance on a multiple-choice credibility judgment test, and in verbal reports with thinking of equal quality. This finding suggests that the exact nature of the request to think aloud is not crucial, and can be adjusted to suit the task and the disposition of particular examinees to verbalize their thoughts. An interviewing procedure consisting of a series of stages progressing from relatively nonleading to relatively leading questions has been used successfully (Norris, 1989). In briefest outline, the procedure follows this order: (a) Inform examinees of the purpose of the interview, namely, to find out what they are thinking as they work
on the problems of the test; (b) give a relatively nonleading direction, such as, "As you do each question tell me all you can about what you are thinking while you are picking your answer"; (c) if examinees do not give complete explanations of their answers, ask, "Could you tell me more about what you were thinking?" and (d) if examinees choose answers but do not respond to the request for verbal reports, then refer to a central piece of information in the item and ask, "Did that information play any part in your thinking?"

Using the verbal reports. Examinees' answer choices compared to the test maker's key is one source of evidence on the quality of their critical thinking. The information in the verbal reports of examinee's thinking is another—relatively independent—source. The verbal reports give rather direct evidence on whether examinees use properly the criteria of credibility judgment that the test is intended to measure. Thus, the correlation of the first measure of the quality of examinees' critical thinking, as indicated by their answer choices, with the second measure, as indicated by their use of the credibility judgment criteria in their verbal reports, gives evidence on how valid it is to infer from examinees' answer choices to their use of the credibility criteria. Details on how both ratings of quality are determined can be found in Norris (1989).

The procedure depends on the two judgments of quality being independent. Suppose, for example, that an item asks whether the statement of a driver involved in an automobile accident or the statement of an eyewitness is more credible, all other things being equal. The driver claims to have used his signal and the eyewitness claims that he did not. Suppose the test maker's keyed answer is that the eyewitness's statement is more credible (because the driver has a conflict of interest). One examinee says the driver is more credible (the wrong answer according to the key), because the driver would have direct knowledge of his own actions. Another examinee says the witness is more credible (the right answer according to the key), because she was watching from the sidelines and would be able to tell whether or not the signal came on. Each examinee's justification is arbitrary: The driver and the witness had equal access to what they reported. So both examinees thought poorly, as revealed by their verbal reports, though one answered according to the key and the other did not. Thus, the judgment of the quality of their thinking based on their verbal reports is independent of the judgment based on their answer choices.

For each trial item, the correlation between the two ratings of critical thinking is computed across the sample of examinees tested. To the degree that the correlation is low for a given item, the use of the credibility judgment criterion cannot be inferred dependably from examinees' answer choices, so the item should be modified or discarded.

The use of standard item statistics, such as item difficulty levels and biserial correlations between item scores and whole test scores, can be used in tandem with data from verbal reports. Sometimes, however, the two sources of information lead to contrary indications. Negative and very low biserials are usually assumed to call items into question. Items that are correctly answerable by most or few examinees would be judged by some evaluators as not suitable for the audience. Such difficulty levels and biserial correlations lead to low estimates of reliability. Some evaluators believe, however, that the traditional psychometric conception of reliability is problematic and that test construction should be guided more strongly by validity considerations (Ennis & Norris, in press). The issue is too complex to treat here, so I shall close with the point that the verbal reporting method urged here is meant to complement as much as possible existing test construction practices.

Making modifications. Based on correlations between the two measures of quality of thinking, as well as the other item statistics, modifications should be made to items, directions, or the key. Modifications are needed if any of the following conditions, among possible others, obtain: (a) Examinees do not understand the task in the intended way; (b) the test, including the answer key, assumes extra-critical-thinking knowledge not held by the intended audience of examinees; (c) examinees think critically according to the evidence in their verbal reports but reach other than the keyed responses—that is, examinees follow the credibility judgment criteria being tested, but, because of some extra-critical-thinking factor, reach answers different from those keyed correct; or (d) examinees think uncritically according to the evidence in their verbal reports but reach the keyed answers—that is, examinees do not follow criteria of critical thinking, but, for sheer extra-critical-thinking reasons, arrive at the keyed responses.

The best modifications to make in view of any of these conditions cannot be determined by a set of rules. Item writing requires creativity and informed judgment, and, though items must be constrained by data, the data underdetermine their exact nature. Generally, many different items can meet the same constraints.

Assessment of the Approach

At the end of the above procedure, we would have direct evidence from verbal reports that examinees who respond in accord with the key tend to use the credibility judgment criteria being tested, and examinees who respond contrary to the key tend not to use the criteria. That is, we would be able to say that in the context chosen, where we know the criteria apply, use of the criteria is what determines differences in scores.

Until the generalizability issue in its various guises is resolved, however, we should not make inferences about examinees' use of the criteria in other contexts. A series of credibility judgment tests with their items cast in various contexts, each designed using the previously described procedure, could be used to help test the various generalizability hypotheses. In this manner, making higher quality tests for restricted ranges of contexts could improve theorizing which in turn could lead to more valid critical thinking testing.

The recommended verbal reporting approach still requires considerable exploration. First, it is not known whether the use of verbal reports of thinking can help with the development of tests of critical thinking dispositions. It seems worth exploring whether interviewing examinees is a way to understand the beliefs and values that underlie the critical thinking abilities they use, but it is not clear how tests of critical thinking dispositions might look.

Second, the methodology as described is not designed for improving critical thinking testing in the context of real-world problems. Typically in such problems, several critical thinking abilities have to be orchestrated, judgments must be made about how to apply ab-
judging credibility—by using verbal reports of examinees’ thinking on items to gain direct evidence on the reasons for their answer choices. Given the reports that most students are adept at locating information but not at using and assessing it (e.g., National Assessment of Educational Progress, 1985), improved tests that enable more accurate monitoring of students’ progress in this area would be a worthy goal. In addition, improved tests that enable more careful study of critical thinking generalizability and critical thinking dispositions would be theoretically and educationally significant.

---

**Note**

This paper is based upon research supported by the Social Sciences and Humanities Research Council of Canada, Grant Nos. 418-81-0781, 410-83-0697, and 410-85-0587. I thank Linda Phillips and two anonymous reviewers for helpful comments on a previous draft.

---

**References**


