4.

Investigating and Assessing Multiple-Choice Critical Thinking Tests

Robert H. Ennis

Stephen Norris and I have long urged (e.g., Norris and Ennis 1989) the following basic steps in the investigation and assessment of a critical thinking test.

1. Make sure that the test is based on a defensible conception of critical thinking that is acceptable to you — and that the test does a reasonable job of covering that conception.
2. Examine the arguments, including your own, regarding the test's validity for students at the level of your students, in a situation like theirs.
3. Take the test yourself and score it with the key or guide to scoring. Assure yourself that the set of answers or the guide is appropriate for the situation.

Although the first and third of these steps are listed separately (to focus on conveniently identifiable actions), they are actually part of the second step, that is, examining the arguments in support of claims about the test's situational validity, the topic of this chapter.

In pursuit of this topic, I shall suggest a structure for appraising a claim regarding the situational validity of a critical thinking test, and apply the structure to a real case: my recent attempt to revise the manual for the **Cornell critical thinking tests** (Ennis, Millman, and Tomko 2005). The suggested structure applies a broad inference-to-best-explanation approach to particular features involved in test validation, and assumes some, but not all, recent stances and insights of leading psychometricians, including Samuel Messick (1989a, 1989b) and members of the Joint Committee on Standards for Educational and Psychological Testing of the American Educational Research Association, the American Psychological Association, and the National Council on Educational Measurement (1999).

Three Contemporary Stances

Validity: A unitary concept

One stance adopted by Messick and the Joint Committee is that validity is a unitary concept; that is, there are not different types of validity, such as criterion validity, predictive validity, content validity, and discriminant validity, but only validity. This stance is not universally accepted, but I shall assume it here, without arguing the point.

Test validity versus validity of inferences from, or interpretations of, test scores

One significant difference between my approach and that of Messick and the Joint Committee concerns the bearer of validity — what it is that can be valid. They hold that inferences from, or interpretations of, test scores, and not tests themselves, are the bearers of validity. This view was endorsed and called the “consensual understanding” by David Frisbie in his 2005 Presidential Address to the National Council on Measurement in Education (Frisbie 2005). I urge something less radical, namely, that a test in a situation (or set of circumstances) is the bearer of (situational) validity. In my paper “Situational Test Validity”, I urge that we define **situational test validity** as follows: A test is a valid test of X in a situation to the extent that it is an adequate measure of X in the situation. It is significant that this definition is situation specific and does not provide “test validity” with a meaning outside of a situation or type of situation.

In this definition, “the situation” can refer to a particular situation in which the specified test has been given or is to be given (for example, the testing of the fifty-two psychology and humanities students in Tom Solon’s (2001) study; or it can refer to a type of situation (such as the testing of lower-division college students under standard conditions). Although both are of interest, the former is of primary interest to a test user. Even if a test is substantially valid in standard situations, what matters most to a user is whether the test was or will be at least substantially valid in the user’s situation. In contrast, people preparing a test manual will be interested in the myriad of situations under which a test might be or has been used, but cannot take account of all of them. For this reason, they are likely to focus on standard situations, or the types of non-standard situations that are most likely to be encountered by test users.

The different kinds of situations imply an ambiguity in the definition of test validity, but it is not a destructive one as long as the person interested in the situational validity of a test is clear about the difference. In my investigation of the situational validity of the Cornell tests, I focused on standard-situation validity, though I was very interested in the particular-situation validity of the tests in the various studies in which they were used.

Non-quantitative appraisal of validity arguments

A third stance adopted by Messick and the Joint Committee, which I infer from what they say about arguments about validity, and with which I agree, is that the strength of a validity argument is not to be stated in numbers (such as 0.82), but in more vague normative terms. The terms they use are “consonant with,” “less well supported,” “scientifically sound… argument,” and “support.” Such words do not invite the attachment of numbers and are not replaceable by them.

In his essay “Validity,” in Robert Linn’s highly regarded third edition of *Educational Measurement* (1989), Messick equates validity with the consonance of evidence for an inference and lesser support for alternative lines of evidence: “To validate an interpretative inference is to ascertain the degree to which multiple lines of evidence are consonant with the inference, while establishing that alternative lines of evidence are less well supported” (Messick 1989a, 13). The Joint Committee has stated that **test validation** is the process of developing a “scientifically sound” validity argument to support an interpretation: “Validation can be viewed as developing a scientifically sound validity argument to support the interpretation of test scores and their relevance to the proposed use” (Joint Committee 1999, 9).

To the terms “consonant with,” “lesser support,” and “scientifically sound,” I would add other words that can be used to express a judgment about degree of support. These words include “fully,” “substantially,” “moderately,” “basically,” “apparently,” “seemingly,” “probably,” “likely,” “for the most part,” “by and large,” “reasonably well,” “sufficiently for the purpose,” “somewhat,” “possibly,” “weakly,” “hardly,” and “minimally.” These words are not replaceable by numbers, but are used to express less precise normative judgments.

The most frequent question I receive about the **Cornell critical thinking tests** (Ennis and Millman 2005a, 2005b) and **The Ennis-Weir Critical Thinking Essay Test** (Ennis and Weir 1985) is “What are the reliability and validity of this test?” This usually means that the questioner thinks that there is a number that can be given for the reliability and the validity of the tests. Reliability indices can be numbers, which are generally correlations (I shall have more to say about this later in this chapter), but if one agrees with the Joint Committee and Samuel Messick, as I do, the degree of situational validity cannot be captured in a number.

Numbers can be attached to correlations with other critical thinking tests and to correlations with other criteria (such as first-year grades in graduate school). Such numbers have been given names like “criterion validity” and “predictive validity.” These names, in accord with the unitary conception of validity assumed earlier, are better expressed as "-related evidence of validity." The first example would then read: "criterion-related evidence of validity." This would make it clear that the numbers sometimes given for validity are evidence for validity, not validity itself.

Best-Explanation Reasoning

Given that the appraisal of situational validity ultimately calls for the construction of an argument, I find it helpful to work from the assumption that such an argument is a best-explanation argument. In this context, the best-explanation argument is an argument in which, very briefly, the hypothesis that a test is valid to a substantial degree in a given situation (or type of situation) is supported by (a) the ability of the hypothesis to best explain, or best contribute to explaining, the observations about the test; and (b) the inability of alternative hypotheses to explain them (roughly what Messick suggested in the quote above).

In more detail, in accord with the broad approach that I have developed,[1] a hypothesis of situational validity is supported roughly to the extent to which, given reasonable assumptions,

1. it can explain (account for) evidence — or help to do so;
2. there is no evidence that is inconsistent with the hypothesis;
3. evidence is inconsistent with alternative explanations of the data;
4. the hypothesis is plausible — it fits with what else we know;
5. realistic and earnest attempts have been made to find counter-evidence and alternative hypotheses;
6. the hypothesis implies new evidence (especially helpful if the new evidence is surprising); and
7. the evidence is well established.

Criteria 4, 5, and 6 overlap at least to some extent with some others, but it is helpful to make them explicit. These three, together with the other criteria, have been topics of discussion and debate for many years, but here I assume them.

Types of Evidence

The best-explanation structure of validation arguments provides broad criteria for making validation judgments. Messick (1989b, 6) has suggested specific types of information that are relevant to these broad criteria. Inspired by his suggestions[2] with some supplementation by me, I propose the following ten (somewhat overlapping) types of evidence that are likely to be relevant when making a judgment about situational validity in regard to a critical thinking test:

1. the rationale upon which the tests are built;
2. the degree to which the tests cover the items in the rationale;
3. reasonable judgments about the acceptability of the keyed answers;
4. internal statistical analyses: item analyses, internal consistency indices (the latter being called "reliability" in psychometric language), and factor analyses;
5. consistency of test results over time for individuals, including test-retest consistency and inter-rater consistency, which are also called "reliability" in psychometric language;
6. appropriate consistency across groups or settings (generalizability);
7. correlations and other relationships between the test and other variables;
8. correlations between the test and other tests of and criteria for critical thinking;
9. results of experimental studies in which teaching critical thinking was attempted, and in which the test was used as an indicator of success; and

10. the extent to which test results fit into our general knowledge, including the contribution the tests have made to our knowledge of the relationship between critical thinking ability and other things.

This list is not intended to include all possible types, but I think it is fairly comprehensive. Each of these types is relevant to one or more of the seven criteria for best-explanation arguments I outlined earlier.

On the basis of my experience revising the Cornell manual, I can testify that a large amount of information must be gathered and interpreted when one makes a validity judgment in accord with the proposed ten types of evidence and the seven criteria for best-explanation arguments. The task is difficult if one is to produce anything approaching a reasonable judgment about validity. This is one of the reasons that validity is often slighted in descriptions of tests. It is much easier and less expensive to present an internal consistency index (by applying a Kuder-Richardson or Cronbach alpha formula to the results of a single administration of a test), which is a number, such as 0.85, and which is misleadingly called “reliability.” More about this later.

An Example: Making a Validity Judgment About the Cornell Critical Thinking Tests

To illustrate the process of making a validity appraisal along the lines just suggested, to exhibit some distinctions and problems, and to show that the process is not an easy one, I shall describe my recent experience with revising the manual (Ennis, Millman, and Tomko 2005).

Level Z is the **Cornell critical thinking test** aimed at gifted and advanced high school students, college students, graduate students, and adults. Level X is aimed at students at middle or secondary levels of education, including 4th or 5th graders under special conditions of administration (Ennis and Millman 2005a, 2005b). My hypothesis is that the two tests I appraised are, to a substantial degree, situationally valid in standard situations, but I shall not here indicate the extent to which I believe the hypothesis to be established. My primary purpose is to present and comment on a process, not to defend a judgment about the Cornell tests.

Anyone trying to develop a picture of the validity of a particular test faces the problem of securing data. Large testing organizations have resources to conduct independent studies, but the cost impinges on their income, so they try to use information from the administration of their tests by other people. In reviewing the Cornell tests, our first problem was to secure data from the use of the tests. We were fortunate that a large number of studies have been done with Cornell Level X and Level Z. For earlier versions — as well as the most recent version — of the manual, we reviewed the *Dissertation Abstracts International* and the *Social Science Citation Index* from 1970 to 2000 to find sources of data. Most sources we located had some usable data. These, combined with several studies we did ourselves and several sent to us voluntarily, resulted in a total of sixty-nine usable studies for Level X and forty-two for Level Z.[3]

I shall refer to some of these studies as I discuss the problems and processes involved in evaluating critical thinking tests in accord with the ten types of evidence in the list above. The first three types in the list come under the heading "content-related evidence of validity."

Evidence types 1-3: A clearly defensible conception of critical thinking and its incorporation in the test

In generating or appraising a test, it is important to have a clear and defensible conception of critical thinking on which the test is based, partly because this will clarify one's hypothesis about the test's situational validity. The presentation of the conception provides the opportunity to decide whether it is close enough to what the test user has in mind, and whether critical thinking so conceptualized is worthwhile (required by the commendatory tone of "critical thinking").

Approaches to critical thinking do vary. Some approaches emphasize the degree to which the argument, presentation, or statement under consideration is persuasive, not whether it is justified. The **Cornell-Illinois conception**,[4] on which the Cornell critical thinking tests are based, is concerned with justification. We might begin with the following brief definition: Critical thinking is reasonable reflective thinking focused on deciding what to believe or do. This definition is too general to provide much guidance in the construction and evaluation of a critical thinking test, but the following more detailed definition can serve as a bridge from the brief definition to an even more detailed specification of abilities and dispositions: Critical thinking is focused, skilled, active, reasonable thinking, incorporating the identification, clarification, and due consideration of the situation, relevant background information, reasons, evidence, and alternatives in deciding what to believe or do.[5]

Based on the brief and the bridging definitions is an elaborate and detailed set of critical thinking abilities and dispositions of critical thinkers. This set can be the basis for a detailed table of specifications for a critical thinking test. The most readily accessible version of this detailed set is the outline of goals for a critical thinking curriculum and its assessment on my academic website (http://faculty.ed.uiuc.edu/rhennis). For a similar presentation in print form, see Ennis (2001); for exemplification and interpretation, see Ennis (1962, 1987, 1991, 1996).

But a clear and defensible conception of critical thinking is not enough. The conception must also be well incorporated in the test. This calls for an examination to determine whether the conception is adequately covered (although complete coverage is unlikely for any test of critical thinking). And whether the keyed answers to test questions are justified. In making this judgment, a prospective user should examine the extent of coverage and take the test, checking the adequacy of the prospective user's answers as well as the answers in the key. The keyed answers for the Cornell tests are defended at the end of the manual, but a prospective user should still take the time to check them.

The Cornell tests do not fully cover the Cornell-Illinois detailed conception, as can be seen in Table 1, which lists most of the main topics included in the detailed conception. One must decide whether the coverage is adequate for one's purposes.

Aspect of Critical Thinking	**Items of Level X (for K-12)**	**Items of Level Z (for UG, Grad, Adult)**
Induction	3-25, 48, 50	17, 26-42
Deduction	52-65, 67-76	1-10, 39-52
Value Judging	Not tested	Not tested
Observation	27-50	22-25
Credibility of Sources	27-50	22-25
Assumption	67-76	43-52
Meaning	Not directly tested	11-21, 43-46
Dispositions	Not directly tested	Not directly tested

Table 1: A rough assignment of test items to aspects of critical thinking[6]

The multiple-choice format has some significant desirable features: multiple-choice tests can be graded easily and cheaply, and can assure coverage of specific aspects of critical thinking. But this also means that some significant aspects of critical thinking mentioned in Table 1 are not tested — value judging and dispositions for both tests and meaning abilities for Level X. It is difficult to have value-judging items because this would probably require the assessment of a test taker's value judgments, which would be unfair. The multiple-choice testing of dispositions would appear to be useful only for situations in which students do not reveal their names to people who matter to them (savvy students are not likely to admit that they are not open minded, for example, even if they are not open minded). And it is difficult to phrase questions designed to test meaning abilities in a way likely to be understood by less sophisticated students.

The creative aspects of critical thinking also tend to be neglected in multiple-choice tests. These include formulating hypotheses, doing the creative parts of planning experiments, formulating definitions, and formulating appropriate questions. These aspects require more open-ended kinds of assessment.

Other limits on multiple-choice testing can be found in attempts to test for skill at best-explanation induction and the judging of credibility. When we draw inductive and credibility conclusions, judge them, and even decide the bearing of evidence upon them, we rely on a vast array of auxiliary assumptions about the way things happen. As in real life, the need for all of these background-belief assumptions exists in a test situation when we ask students to make a commitment to some view that students might not share.

A second problem arises because a less sophisticated person is sometimes justified in calling true something that a more sophisticated person would justifiably call only probably true. In the same circumstances, a very sophisticated person might justifiably judge that there is insufficient evidence for either position (problems Groarke raises in his chapter in this volume with respect to certain multiple-choice questions on the *California Critical Thinking Skills Test*). These problems can be reduced in best-explanation induction test items by asking for merely the direction of evidential support, if it has a direction, rather than whether the conclusion is true, probably true, etc. With credibility test items, one can ask which of two statements is more credible, if either is, instead of asking whether a statement is credible. This again avoids the requirement that one make an absolute judgment.

The first problem with best-explanation induction and the judging of credibility is somewhat more difficult to handle, because different people bring different auxiliary assumptions to bear on decisions of this sort. Though not always a solution, the most reasonable approach calls for auxiliary assumptions on which most people will agree. For example, we believe most people would agree on the following auxiliary assumption of Item 1 of Level X: “If a hut is not lived in or used, a layer of dust will probably develop.” But we are not certain that *all* test takers would agree on even this auxiliary assumption, and do not want to penalize them for holding a different belief about the way the world works. Accordingly, we have provisionally adopted a stance that deems as indicative of mastery any induction or credibility section score with a greater than 85 percent agreement with the answer key.

These content problems must be faced in making a situational-validity judgment.

Evidence types 4 and 5: Internal consistency and consistency over time (“reliability”)

Some internal consistency is desirable because a test should hang together in some reasonable way if it is to be named by a single noun or noun phrase, such as “critical thinking.” Standard measures of **internal consistency** are the extent to which students who do well on the total test do well on a particular item (item discrimination), and (roughly) the average correlation of each item with every other item. The latter is what we get with the Kuder-Richardson and Cronbach alpha formulas, which are the indices most frequently used and reported under the label “**reliability**.”

Calling these indices “reliability indices” is unfortunate (Ennis 2000) because they indicate only internal consistency, not what is ordinarily meant by “reliability” (a combination of consistency and accuracy). According to the psychometric concept of *reliability* (which is only consistency, whether internal or not), a bathroom scale that consistently reads 15 pounds low is totally reliable, as is a compass that consistently reads 180 degrees off (that is, reads just the opposite of what it should read). This is a serious problem because, when information is called “reliable,” many test users think they are being given test-validity information.

Internal-consistency psychometric reliability is especially attractive to test makers. The numbers run higher than validity-related numbers and they are inexpensive to secure. One has only to run a computer program on the item scores obtained in one administration of a test to get an internal-consistency index. Consequently, test makers can get inexpensive and misleading indicators that are indicators only of internal consistency, but advertise them as “reliability” indices. Inevitably there is pressure on test makers to increase the internal-consistency indices.

One way for test makers to increase internal consistency is simply to lengthen the test by adding more, similar items. Another way is to discard any items that do not correlate well with the total score, that is, those with low item discrimination (also a misleading label, unless the test is **uni-dimensional**). This increases the correlations items have with each other and thus internal consistency, but also increases the uni-dimensionality of the test.

But critical thinking is not uni-dimensional, as can be seen by looking at the wide variety of aspects associated with it (as Johnson argues in his chapter in this volume). For example, in Cornell Level Z, deduction, meaning, fallacies, observation, credibility of sources, hypothesis testing, planning experiments, definition, and assumption identification are all assessed.[7]

Empirical support for the **multi-dimensionality** of critical thinking appears in the Level Z manual (Ennis, Millman, and Tomko 2005; from Mines 1980). Part-score “reliabilities” for Level Z ran almost as high as the total-score “reliability.” That is, 0.76, 0.66, 0.60, 0.55, 0.72, 0.65, and 0.65 are about as high as 0.76, the “reliability” for the total score (N=40 graduate students at the University of Iowa). Adjusting these part-score “reliabilities” (using the Spearman-Brown formula) for the lengthening of each part to 52 items (the actual number of items in the total test), these part-score “reliabilities” become 0.94, 0.90, 0.95, 0.83, 0.97, 0.96, and 0.94, which are considerably higher than the 0.76 for the full test. This strongly suggests multi-dimensionality and justifies not expecting internal-consistency (“reliability”) indices of over 0.80 on comprehensive critical thinking tests. A very high internal-consistency index (e.g., 0.92, or 0.95) would be undesirable. Admittedly, this reasoning involved stretching the Spearman-Brown formula beyond its original intent, but the results are still rather striking.

A second and more defensible kind of internal consistency for multi-dimensional critical thinking tests is “split-half” consistency.

In this case, a test is split in half (typically into odd items and even items, sometimes into equal-length sets of items judged comparable), the two halves are correlated with each other, and the correlation is adjusted upward (by the Spearman-Brown formula) to compensate for each half's being shorter than the full test. Split-half consistency is more defensible than some of the other measures of consistency because sums of composites are correlated with sums of fairly comparable composites (assuming a roughly equal number of items from each part of the test), instead of each item being correlated with every other item. Computing such measures is more troubling than the Kuder-Richardson and Cronbach alpha internal-consistency estimates, however, and it is still misleading to call split-half internal consistency "reliability," because it provides a measure of consistency, not situational validity.

Another type of consistency is **test-retest consistency**. It is not vulnerable to the complaint that it unduly penalizes a test for multi-dimensionality. But it, too, is wrongly called "reliability." It is (only) a measure of consistency from one administration of a test to the next, and does not show that the test is assessing what it claims to be in the situation. Test-retest consistency is investigated less often. Many things can happen from one administration of a text to the next and this may interfere with a consistency measure. Even without this complication, test-retest consistency is generally avoided because the required two administrations (reasonably separated in time) of the same test to the same population are generally much more trouble than one administration.

Inter-rater consistency is important for tests, typically essay tests, that must be graded according to some rubric or criterion. Again, consistency is not validity though it is more likely to tend in that direction if the graders are familiar with the goals and their meaning, and if they are competent. Inter-rater consistency, however, does not apply to multiple-choice critical thinking tests, the topic of this chapter.

For both Cornell tests combined, we have twenty-six examples of Kuder-Richardson internal consistencies, as contrasted with fourteen examples of the split-half type of internal consistency, and two examples of test-retest consistency. Variation among groups and settings is expectable, but simple arithmetic means give a good indication of central tendencies.[8] For Level X the simple mean for Kuder-Richardsons is 0.79, and for split halves it is 0.83; for Level Z, it is 0.67 for Kuder-Richardsons and 0.67 for split halves. For identifiable graduate students on Level Z, the split-half internal consistencies averaged 0.78, and the only Level Z Kuder-Richardson I found is 0.76. This suggests that Level Z is more internally consistent for more sophisticated students than it is for less sophisticated students. The test-retest consistencies were obtained for Level Z only, and averaged 0.79. Results like these are quite acceptable, if the multi-dimensionality thesis is acceptable. These “reliabilities” are not as high as those in good uni-dimensional tests, such as the verbal, quantitative, and analytic parts of the former Graduate Record Examination (GRE) general test, which are listed at 0.92, 0.92, and 0.88 respectively (GRE Board 1995-6, 30).

The simple mean of the item discrimination indices is 0.24 (N=6) for Level X and 0.22 (N=5) for Level Z. These are reasonable, especially for multi-dimensional tests. Item discrimination is a type of internal consistency that is not called “reliability,” and not one that yields anything like the high numbers of Kuder-Richardsons.

The Kuder-Richardsons and other internal-consistency results are roughly explainable by the tests’ being multi-dimensional and the Level Zs’ being aimed at more sophisticated students. As such, they are quite adequate, though not as high as the ones in uni-dimensional tests, for example, the verbal and quantitative parts of the GRE, which run approximately 0.92. But note that the GRE program did not combine three components — verbal, quantitative, and analytic — to compute internal-consistency estimates. Combining them would produce a multi-dimensional test and lower the internal-consistency index.

It is sometimes held that psychometric reliability is a necessary condition for validity. This is generally true for test-retest and split-half consistency. A test with inconsistent retest results raises the question “Which is right: the test or the retest?” And a test with inconsistent, supposedly comparable halves would seem odd. But for intercorrelation internal-consistency indices, the claim that consistency is a necessary condition for validity is an exaggeration because of the multi-dimensionality possibility, although at least some internal consistency is generally desirable for a test named by a noun.

In sum, when judging consistency one must carefully consider the type of consistency measure used and interpret it accordingly. It is important to compare critical thinking tests using the same type of consistency, and, if comparing internal consistencies, to consider whether the tests attempt to assess only one or a few similar aspects of critical thinking, or attempt to assess a more comprehensive conception of critical thinking. In all considerations of consistency, it is important to be wary of treating consistency (psychometric reliability) as validity.

Evidence types 6 and 7: Relations with other factors, and appropriate generalization

One can generalize from the twenty studies that checked for gender differences using either Cornell Level X or Level Z. There seems, in general, to be no difference in critical thinking ability between mature males and females, assuming that these tests were valid in the situation of their use, although there was some evidence for the superiority of females among younger students. Using Level X, there was occasionally a leaning toward a conclusion that girls were better critical thinkers, but with Level Z, there was no indication of superiority of one gender over the other. The slight difference between the tests could result from the fact that Level Z is given to older students. It is possible that girls are a bit more advanced in critical thinking in grades four to twelve, as they are in many mental activities, and that boys catch up in college and above.

The results for gender seem consistent across groups. The values of the gender variable are clearly identifiable, and we have reason to expect that males and females who are tested together represent the same level of critical thinking ability within their gender groups (that is, that the males were roughly in the same male percentile range as the female percentile range of the females to whom they are compared). The results are explainable by these factors, by the combination of a set of what I believe to be reasonable beliefs about male/ female critical thinking levels, and by the hypothesis that the tests were valid in the situation of their use.

In contrast, one would expect less consistency in relationships between test results and grades given by an instructor, because there is considerable variation in the types of prowess rewarded by grades in institutions in the United States. With both Level X and Level Z we found relationship to grade point averages to vary considerably. The greatest disparity was for Level Z, its correlations ranging from –0.02 to +0.60. The 0.60 value was obtained at Cornell University. This is in keeping with my experience there, which leads me to believe that critical thinking is commonly taught and rewarded at Cornell. It is different, however, from experiences in other situations. But even with the obtained variation, the central tendency in the studies surveyed is a moderate relationship with grades.

The lesson here is that complete generalizability is not always to be expected. What should be expected depends on the situation, including the factors related to critical thinking ability, and whether generalizability, or lack of it, can reasonably be explained. Generalizability is more expectable for gender than for grades. For grades, less generalizability is expectable.

Other areas of seemingly moderate consistency were also evident. They included improvement in critical thinking across grade levels; negative correlations with dogmatism; low positive correlations with socio-economic status, independence, and first-year grades in graduate school (the latter being about the same as those obtained with the Graduate Record Examination and the Miller Analogies Test; see Linn 1982); and moderate correlations with IQAA (IQ and Academic Admissions tests) and grades, though there were wide variations for grades, as I pointed out earlier. All of these findings are explainable by the hypothesis that the Cornell tests are situationally valid, together with other plausible assumptions, for example, the assumption that critical thinkers are not dogmatic.

Evidence type 8: Other critical thinking tests

The correlations between the Cornell tests and other critical thinking tests, especially the **Watson Glaser test**, are high to moderate, and are explainable by the situational validity hypothesis, taken in conjunction with the assumption that the tests assess some things in common but also differ somewhat.[9] It is unfortunate that there are not more data for correlations with other critical thinking tests, but they are difficult to obtain, partly because students and teachers resist testing that is done solely for the sake of research. Because it is desirable that an argument for the situational validity of a critical thinking test include correlations with other critical thinking tests, the situational-validity hypothesis receives less support or challenge here than it should. Ideally, for the hypothesis, there would be more correlational studies with other critical thinking tests, producing fairly high to high correlations, depending on the nature of the tests.

Evidence type 9: Experimental studies of teaching

Suppose that, in a teaching experiment, the experimental group improved significantly more than the control group. If critical thinking had been taught — and taught well — to the experimental group only and the experiment had otherwise been run well, then the hypothesis that the test was a situationally valid test of critical thinking is supported. This is because, together, the hypothesis and the two conditional clauses above roughly explain the results. The hypothesis gets further support if the two conditional clauses are established, and the explanation of the results is not plausibly completed other than by the situational-validity hypothesis.

A third type of support can come from a situation in which critical thinking is not taught (even if the investigator thought it was, or might have been), the experiment is otherwise run well, and the experimental group does no better on the test than the control group. The situational-validity hypothesis would help explain the lack of difference between the experimental and control groups. So, in this type of case, negative results would also support the hypothesis.

The above reasoning is schematic, but it shows how **best-explanation reasoning** can guide our thinking about the relevance of experimental results. There are other possible combinations of the factors involved, but these three exhibit a general strategy when there is a control group. When there is no control group, but only a test-retest situation, support provided by positive and negative results is generally weaker because there is more opportunity for other possible explanations of the results.

The application of the above type of schematic thinking is difficult because each case is unique — with many details in doubt, even for the investigator. Nevertheless, experimental evidence is relevant, even though claims about its relevance must usually be qualified by words like “probably,” “possibly,” and “it seems that…”.

From the twenty-seven experimental reports using Cornell Level X that we found, it seems that all but one provided support for the hypothesis. Some of the experiments seemed bound to fail because of the nature of the experimental variables (some of which I think were mistakenly called “critical thinking”), and they did fail to yield a significant difference. Others seemed likely to succeed because a reasonable conception of critical thinking was used, critical thinking principles were made explicit, and probably sufficient time was devoted to the task. In all but one case these experiments did succeed. These results lead me to say that the situational-validity hypothesis for Level X is substantially supported.

In the Level Z experiments with college students, the desirable conditions for learning critical thinking seemed to be present and statistically significant results were obtained in all four experiments we found — with respectable Cohen's *d*'s of 1.1, 1.5, and 0.6 (Cohen 1992) in those experiments that produced this statistic (Solon 2001, 03). The situational-validity hypothesis for Level Z, together with additional assumptions, explains these consistently favourable results. But more data are needed.

Evidence type 10: Contributions to knowledge

Tom Solon (2007), the investigator in some above-reported experiments using Level Z, asserts that his experimental class in which he infused critical thinking in psychology instruction did as well in psychology as the one in which the infusion did not occur. In the other two studies I found that investigated the matter, subject-matter comprehension did not suffer. This is not difficult to understand because the involvement occasioned by critically thinking about the subject matter could easily compensate for the reduced time spent on standard subject-matter instruction. In this context this result satisfies the sixth best-explanation criterion, "the hypothesis [helps imply] new evidence, especially if the new evidence is surprising," and constitutes the tenth evidence type, "contributions the tests have made to knowledge."

Other contributions to the sixth best-explanation criterion and the contribution-to-knowledge type of evidence are the findings about gender, grades, socio-economic status, independence, dogmatism, IQAA, and general improvement in critical thinking across grade levels. These findings are explained by the situational-validity hypothesis and a set of plausible assumptions. As in the case of experiments, more data would, of course, be helpful.

Summary and Comment

In this chapter, I propose a program for investigating and assessing multiple-choice critical thinking tests. The program assumes a focus on the test and the testing situation rather than on inferences from, and interpretations of, test scores. In concurrence with psychometric lore.

I have assumed that numbers are not a good way to try to indicate the extent of validity, and have assumed a unitary conception of validity.

It is helpful to view test-validity claims as **inference-to-best explanation** hypotheses which can be assessed on the basis of seven criteria. A hypothesis of situational validity is supported roughly to the extent to which, given reasonable assumptions, (1) it can explain (account for) evidence — or help to do so; (2) there is no evidence that is inconsistent with the hypothesis; (3) evidence is inconsistent with alternative explanations of the data; (4) the hypothesis is plausible — it fits with what else we know; (5) realistic and earnest attempts have been made to find counter-evidence and alternative hypotheses; (6) the hypothesis implies new evidence (especially helpful if the new evidence is surprising); and (7) the evidence is well established.

Ten categories of information inspired by a list by Messick (1989b) particularize the best-explanation approach for this context:

1. the rationale upon which the tests are built;
2. the degree to which the tests cover the items in the rationale;
3. reasonable judgments about the acceptability of the keyed answers;
4. internal statistical analyses — item analyses, internal consistency indices (called “reliability” in psychometrics), and factor analyses;
5. consistency of test results over time, including test-retest consistency and inter-rater consistency (also called “reliability” in psychometrics);
6. appropriate consistency across groups or settings (generalizability);
7. correlations and other relationships between the test and

other variables;

8. correlations between the test and other tests of and criteria for critical thinking;
9. results of experimental studies in which teaching critical thinking (or something else) was attempted, and in which the test was used as an indicator of success; and
10. the extent to which test results fit into our general knowledge, including the contribution the tests have made to our knowledge of the relationship between critical thinking ability and other things.

By looking at these categories in the case of the Cornell critical thinking tests, I have tried to illustrate the complexities involved in making a reasonable validity decision about critical thinking tests, the difficulty of obtaining firm and clear results in critical thinking research, and the need for attending to many features of the situations in which the data were (or might be) obtained. The resulting challenge may, in part, explain test makers' heavy reliance on psychometric reliability, which is fairly easily determined and a misleading name for consistency. In examining consistency, it is important to be aware of the kind of consistency that is claimed for any test. Different types of psychometric-reliability consistency vary in their import, partly because tests vary in their degree of uni-dimensionality and partly because different factors can be checked for consistency.

The desirability of the generalizability of relationships depends on the factor which is in question. For instance, considerably less consistency in relation to subject-matter grades than in relation to gender is to be expected for critical thinking.

As a by-product of this investigation and assessment of the situational validity of the Cornell tests, some of the more interesting results of a review of the literature using the Cornell tests are (1) the genders are essentially equal in critical thinking ability, given mature students (though among less mature students, girls might have an edge); (2) there is a great deal of variation in the sorts of activities that people evaluate for their efficacy in promoting critical thinking; (3) critical thinking can be taught; (4) infusing critical thinking into subject-matter instruction does not appear to interfere with subject-matter learning; (5) critical thinking is a multi-dimensional concept; (6) critical thinking is negatively related to dogmatism; and (7) critical thinking is positively related to independence, socio-economic status, IQAA tests, subject-matter grades (though there is variation here, presumably attributable to institutional and classroom variation in what is valued and taught), and (using Level Z only) first-year grades in graduate school.

These results are subject to further investigation and depend on the situational validity of the tests used to produce them. This reflects the standard bootstrap situation in science: these results are part of the support for the situational-validity hypothesis, and the hypothesis is part of the support for the acceptability of the results.

References

Cohen, J. 1992. A power primer. *Psychological Bulletin* 112(1): 155-9.

Ennis, R. 2001. Goals for a critical thinking curriculum and its assessment. In *Developing minds*, 3d ed., ed. A. Costa, 44-6. Alexandria, VA: ASCD.

____ . 2000. Test reliability: A practical exemplification of ordinary language philosophy. *Philosophy of education* 1999. Champaign, IL: Philosophy of Education Society.

____. 1996. *Critical thinking*. Upper Saddle River, NJ: Prentice-Hall.

____. 1991. Critical thinking: A streamlined conception. *Teaching Philosophy* 14(1): 5-25.

____. 1987. A taxonomy of critical thinking dispositions and abilities. In *Teaching thinking skills: Theory and practice*, ed. J. Baron and R. Sternberg, 9-26. New York: W.H. Freeman.

____. 1968. Enumerative induction and best explanation. *The Journal of Philosophy* 65(18): 523-9.

____. 1962. A concept of critical thinking. *Harvard Educational Review* 32: 81-111.

____. 1958. An appraisal of the Watson-Glaser critical thinking appraisal. *Journal of Educational Research* 52: 155-8.

Ennis, R., and J. Millman. 2005a. *Cornell critical thinking test, Level X*. Seaside, CA: The Critical Thinking Company.

Ennis, R. 2005b. *Cornell critical thinking test, Level Z*. Seaside, CA: The Critical Thinking Company.

Ennis, R., I Millman, and T. Tomko. 2005. *Cornell critical thinking tests Level X & Level Z manual*, 4th ed. Seaside, CA: The Critical Thinking Co.

Ennis, R., and E. Weir. 1985. *The Ennis-Weir critical thinking essay test*. Pacific Grove, CA: Midwest Publications.

Frisbie, D. 2005. Measurement 101: Some fundamentals revisited. *Educational Measurement: Issues and Practice* 24(3): 21-8.

Graduate Record Examination Board. 1995-96. *Guide to the graduate record examination programs*. Princeton, NJ: Educational Testing Service.

Harman, G. 1973. *Thought*. Princeton, NJ: Princeton University Press.

____. 1968. Enumerative induction and inference to best explanation. *The Journal of Philosophy* 65(18): 529-33.

____. 1965. The inference to the best explanation. *Philosophical Review* 74(1): 88-95.

Joint Committee on Standards for Educational and Psychological Testing of American Educational Research Association, American Psychological Association, and National Council on Educational Measurement. 1999. *Standards for educational and psychological tests*. Washington, DC: American Educational Research Association.

Linn, R. 1982. Ability testing: Individual differences, prediction and differential prediction. In *Ability testing: Uses, consequences, and controversies* (Part II: Documentation section), ed. A. Wigdor and W. Garner, 335-88. Washington DC: National Academy Press.

Linn, R., ed. 1989. *Educational measurement,* 3d ed. New York: Macmillan.

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

Messick, S. 1989a. Validity. In *Educational Measurement*, 3d ed., ed. R. Linn, 13-103. New York: Macmillan.

____. 1989b. Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher* 18(2): 5-11.

Mines, R. 1980. Levels of intellectual development and associated critical thinking skills in young adults. *Dissertation Abstracts International* 41: 1495A.

Norris, S., and R. Ennis. 1989. *Evaluating critical thinking*. Pacific Grove, CA: Midwest Publications.

Solon, T. 2007. Generic critical thinking infusion and course content learning in introductory psychology. *Journal of Instructional Psychology* 34(2): 95-109.

____. 2003. Teaching critical thinking: The more, the better. *The Community College Enterprise* 9(2): 25-8.

____. 2001. Improving critical thinking in an introductory psychology course. *Michigan Community College Journal* 7(2): 73-80.

Notes

1. Although somewhat similar in spirit to the best-explanation-inference approach advocated by Gilbert Harman (1973), my approach does not treat enumerative induction as a special case of best-explanation inference (Harman 1965, 1968; Ennis 1968), and adds some popular features.
2. Omitting his controversial value-implication and social-consequences criteria.
3. Both locating and reviewing these studies were difficult, and we are deeply indebted to the University of Illinois Library.
4. I call it the "[pb_glossary id="364"]Cornell-Illinois conception[/pb_glossary]" because it was conceived and developed while I was at these two universities, where I had much help from colleagues, students, and administrators. John McPeck called the first readily available statement of this conception (Ennis 1962) "the prevailing view of the concept of critical thinking" (see Chapter 3 in McPeck 1981).
5. I am indebted to Michael Scriven for some content of this bridging definition.
6. Reproduced with permission from the Critical Thinking Company (www.CriticalThinking.com).
7. See the "Outline of goals..." on my academic website (http://faculty.ed.uiuc.edu/rhennis) for a more complete list.
8. For simplicity, I used ordinary averages rather than go through Fisher's *z* transformations because it makes so little difference in this situation.
9. In the past, I argued (Ennis 1958) that the then-current version of the Watson-Glaser test had significant problems. Most of those concerns still hold.