

Nationwide Testing of Critical Thinking for Higher Education: Vigilance Required

ROBERT H. ENNIS

University of Illinois Urbana-Champaign

Abstract: The Spellings Commission recommends widespread critical-thinking testing to help determine the “value added” by higher education institutions—with the data banked and made available (“transparent”) in order to enable parents, students, and policy makers to compare institutions and hold them accountable. Because of the likely and desirable promotion of critical thinking that would result from the Commission’s program, I recommend cooperation by critical-thinking faculty and administrators, but only if there is much less comparability and considerably deeper transparency of the tests and their justification than the Commission recommends, and only if vigilance in handling the many problems and dangers elaborated herein is successful.

The nineteen-member Commission on the Future of Higher Education, sometimes called “The Spellings Commission,”¹ has urged extensive critical-thinking testing for college students. They call it “value added” testing because students generally are to be tested at the beginning and end of their college careers to see how much they have improved. The Commission wants the testing results to be transparent and numerically comparable,² so that consumers and policy makers can compare institutions and hold them accountable. Because of the transparency and comparability requirements, I shall argue that two possible consequences are 1) the use of only one or a few nationwide tests and 2) the existence of strong temptations for officials and others to manipulate the situations and the data to make their institutions look good.

Although the Chair of the Commission, Charles Miller, publicly rejects the one-or-a-few-tests consequence (Chaker 2006: A12), I think both consequences are likely if we have both transparency and comparability, exacerbating the problems and dangers that will inevitably be present in such a high-stakes testing situation. The first consequence will contribute to the second consequence, but, even if we avoid the

first, we will still have the second consequence to a great extent, if we have both transparency and comparability.

However, because critical thinking is so important, and widespread critical-thinking testing could be very helpful in promoting it, if done well, I shall here urge substantial cooperation by critical-thinking faculty and administrators.³ More specifically, I shall urge them to support a nationwide critical-thinking testing effort, though seeking much less comparability and much more transparency than the Commission recommends, and to be proactively vigilant about transparency and comparability as well as about the many problems and dangers that I shall presently depict.

In what follows, I plan to present a brief summary of the report and some comments made by Chair Miller, who is a business executive and investor who “was head of the Regents of the University of Texas when they directed the University’s nine campuses to use standardized tests to prove that students were learning” (Arenson 2006: 1A). Among other things, I shall show that the Commission emphasizes critical thinking as a “student outcome.” Then I will summarize the comments made about some of the issues by participants in an electronic discussion on the listserv of the Association for Informal Logic and Critical Thinking (AILACT) in February 2007. Subsequently, I shall offer to critical-thinking faculty and administrators⁴—as well as policy makers and higher-education clients—elaboration and warnings of some problems and dangers that they would face, in the hope that their vigilance, knowledge, and understanding will help them to handle these problems and dangers. Lastly, I shall offer my primarily positive recommendations about the critical thinking–assessment aspects of the report.

Although I here discuss assessment problems and dangers in the context of the Spellings Commission report, these problems and dangers are not unique to this report. They generally accompany assessment pressures, whether exerted by other nations or governments, including states and provinces, or by accrediting agencies.

A Brief Summary of the Report

The newspaper accounts I have seen have emphasized the accountability aspects of the report. The following quote from the report’s summary gives an idea of the nature of the report’s focus on accountability, accountability being an increasingly emphasized idea in U.S. K–12 education in the past twenty-five years, now being extended to higher education:

We believe that improved *accountability* is vital to ensuring the success of all the other reforms we propose: Colleges and universities must become more *transparent* about cost, price, and *student success outcomes*, and *must will-*

ingly share this information with students and families. Student achievement, which is inextricably connected to institutional success, must be measured by institutions on a “value-added” basis that takes into account students’ academic baselines when assessing their results. This information should be made available to students, and reported publicly in aggregate form to provide consumers and policymakers an accessible, understandable way to measure the relative effectiveness of different colleges and universities. (USDOE 2006: 4, italics added)

Other Important Aspects

Some other important aspects of the report (on which I will not focus here) are the affordability of higher education; financial aid programs; access to higher education; part time students; complaints about secondary schools’ not providing colleges with students of a caliber that higher educators desire, resulting in remedial courses; and emphasis on innovation, which becomes in this report, an emphasis on a number of important subject matter areas that some people apparently feel might otherwise be neglected by the testing and reporting programs envisioned in the above accountability statement. Areas mentioned include science, mathematics, technology, engineering, teaching, management, medicine, and foreign languages.

Student Outcomes, Including Critical Thinking

What are the proposed “student success outcomes” with which the accountability statement is concerned? Clearly critical thinking and literacy are central. In his remarks quoted in February 2006, Chair Miller mentioned only “writing, critical thinking, and problem solving” and “analytical reasoning” (Arenson 2006: 1A). The report itself mentions “reading, writing, and thinking” (USDOE 2006: x), and “critical thinking, writing, and problem solving” (USDOE 2006: 3). The two tests that the report recommends as providing “quality assessment data” (USDOE 2006: 24) claim respectively to measure “critical thinking, analytic reasoning, problem solving, and written communication (Council for Aid to Education: 3) and “critical thinking, reading, writing, and mathematics” (Educational Testing Service 2007: 1). These two tests are named, respectively, *Collegiate Learning Assessment* (CLA) and *Measure of Academic Proficiency and Progress* (MAPP). Although critical thinking, literacy, and perhaps mathematics seem to be the proposed basic student success outcomes (assuming that problem solving and analytical reasoning are closely related to, or part of, critical thinking), I shall here limit my discussion to critical thinking.

*Summary of the Opinions of Some Members of the
Association for Informal Logic and Critical Thinking
(AILACT)*

In February 2007, I presented the national critical thinking–assessment topic to the participants in the listserv, AILACT-D, which is the electronic discussion arena for the Association for Informal Logic and Critical Thinking (AILACT). To my knowledge, this is the only national or international professional association devoted to the field of critical thinking. There was a vigorous discussion consisting of fifty-one e-mail messages. Among the discussants, there was a strong rejection of a national government's requiring one or just a few critical-thinking tests. Having a single nationwide critical-thinking test was especially strongly condemned on the ground that it would inevitably be "dumbed down" and politicized. There seemed to be strong support for locally controlled critical-thinking testing, so long as the testing is supervised and controlled by the total local institution—that is, much more broadly controlled than by each instructor for her or his own class, or by any single department. This testing, it was felt, might well be part of the accreditation process. In addition, the conception of critical thinking and the nature of the associated test under consideration were also deemed crucial by AILACT discussants. The recommendations I shall later present are in accord with this sense of the discussion.

Problems and Dangers

In this section, I shall first consider two basic concerns in all assessment, test validity and reliability, starting with one important element of validity of a critical-thinking test, the conception of critical thinking on which a critical-thinking test is based (or which it is alleged to be assessing). I shall then examine in general terms some other problems and dangers associated with the Commission's approach to critical-thinking accountability.

The Conception of Critical Thinking on which a Test Is Based

There is a vast literature on the nature of critical thinking. Approaches vary in accord with the breadth of coverage, amount of detail provided, and the assumed purpose when thinking critically. According to the epistemic approach, which I believe best fits everyday usage of the term 'critical thinking,' the purpose when thinking critically is to find the truth, or the most reasonable approximation thereof—to "get it right" so to speak. This does not require that critical thinking always result in the truth, rather that truth (or its most reasonable approximation, or getting it right) be the goal. Accordingly, the dispositions and abilities

of a critical thinker, roughly speaking, would be those that promote the pursuit of truth, and it is these dispositions and abilities that would be assessed in a fully comprehensive critical-thinking test.

Detailed epistemic approaches to a conception of critical thinking include two Delphi (consensus-seeking) approaches, one led by Peter Facione (1990), and one by the National Center for Educational Statistics (1995). The former is based on a survey of mostly faculty, about half philosophers, and the latter is based on a survey of faculty, employers, and policymakers. A third detailed, epistemic approach, the Cornell-Illinois approach (Ennis 1962, 1987, 1991, 2002a, 2002b), includes criteria to be used by a critical thinker in making judgments.

Less detailed epistemic approaches include those of Alec Fisher and Michael Scriven (1997), Ralph Johnson (1996), John McPeck (1981), Richard Paul (1993), and Harvey Siegel (1988). Ralph Johnson and, to some extent, Richard Paul have gone beyond straight epistemic views to include dialectical features, such as taking into account in advance the objections or concerns of others. These features are needed not only for the pursuit of truth, but also for the sake of fairness and consideration of others—when one thinks critically in formulating and examining real arguments and decisions in context.

Although there has been much discussion of the above conceptions, they generally have the same basic concern: to seek the truth—to try to get it right. To save space for purposes of this presentation, I will not further elaborate them.

The epistemic view should be distinguished from views that the purpose of critical thinking is effective persuasion, resolution of disputes, relativism (agreement with some group or culture), intractable skepticism, cynicism, or routine subject-matter knowledge, approaches that I think do not fit everyday usage of the term 'critical thinking,' and that, based on my reading of the document, I do not think the Commission had in mind. Although learning how to be persuasive is important in a number of circumstances, as are dispute resolution, conforming to a given group or culture, and routine subject-matter knowledge, none focuses on the general traits and knowledge involved in the non-routine seeking of the truth. I do not think that the employees whose complaints about their employees' deficiency in critical thinking were mentioned in the Spellings Commission report (USDOE 2006: 3) were complaining about the employees' inability to persuade, to resolve disputes, to conform, or to be intractably skeptical; or their lack of routine subject-matter knowledge.⁵

However, deep subject-matter knowledge and understanding that includes the abilities and dispositions to think critically within that subject (of the sort expected to be exhibited in doctoral dissertations and by professional subject-matter researchers) would thereby be in-

cluded in epistemic subject-specific critical thinking. Unfortunately, professed exclusive subject-specific critical-thinking proponents (*e.g.*, McPeck 1981) do not seek transfer of these abilities and dispositions to other areas. (If they did seek such transfer, then the critical-thinking abilities and dispositions they endorse would be general, rather than only subject-specific, as they claim.) Most people hold that transfer to relevant areas is desirable for critical-thinking abilities and dispositions. This topic is more fully discussed in Ennis 1989, 1990, and 1992.

In what I have just said, I have not denied the importance (in critical thinking) of familiarity with, and knowledge of, the area or topic in which the thinking occurs. On the contrary, I urge approaching any issue with sufficient Sensitivity, Experience, Background Knowledge, and Understanding of the Situation (SEBKUS) (Ennis 2004).

A danger in not ascertaining and evaluating the underlying conception of critical thinking in a test is that we will unthinkingly accept what might be called a "critical-thinking test" just because it is so labeled. We must consider whether it is based on the everyday (epistemic) concept of critical thinking.

Situational Validity

The most important feature of an educational test is its situational validity, which is different from formal-logic, deductive validity, and from psychometric reliability. Roughly speaking, I define the 'situational validity' of an educational test as the extent to which it assesses what it is supposed to be assessing in a given situation or type of situation. The conception of critical thinking on which a critical-thinking test is based is a key aspect of the situational validity of a test, but there are other important aspects as well.

The above definition of situational validity has some similarity to the conception of test validity that was popular fifty years ago, which did not mention the situation and is more conservative than the one that many psychometric leaders recommend these days.⁶ The main reason for introducing the word 'situational' is that a given test might validly assess a trait in one situation, but not another (*e.g.*, a critical-thinking test written in English for fluent speakers of English, used with students for whom English is not the first language), making test validity relative to the situation or type of situation.

A problem for test makers is that a strong argument for the situational validity of a test, even in standard situations, is difficult to develop. In my view, situational test-validity claims are hypotheses that, roughly speaking, get their support from, among other things, their ability to explain the evidence, and from the inconsistency of alternative hypotheses with evidence (see Ennis, in press; inspired by Harman 1973). Not only are arguments for validity difficult to

develop, but the associated correlational evidence tends to consist of much lower correlation numbers than the evidence for psychometric reliability (that is, consistency—more later). For example, correlations of admissions tests with first year grades, which are commonly mentioned in discussions of the validity of admissions tests (under the label 'predictive validity,' or 'predictive evidence of validity') tend to range between .20 and .40 (Linn 1982), compared with around .90 for psychometric reliability (1.00 being perfect correlation). I use admissions tests here to exemplify the point because they are heavily used, thoroughly studied, high-stakes tests, and many people are familiar with them. The numbers vary with the test, but test validity-indicating numbers tend to be considerably lower than numbers that indicate psychometric reliability.

Furthermore, especially for multiple-choice tests, reliability evidence is much easier to acquire than validity evidence, since it is often provided by a single administration of a test. The resulting danger is that there might be insufficient attention to validity in the promotional materials, and possibly also in the development of the critical-thinking tests being considered for assessing critical-thinking value added. Hence I recommend that critical-thinking faculty and administrators never be satisfied with reliability figures alone, but be sure to ask for, and appraise, the argument for the validity of the test in their situation or in situations like the one being faced. The argument should rest on, among other things, an acceptable conception of critical thinking and the relation between the test and the conception. See Messick 1989a: 6 and Ennis (in press) for a number of factors that are relevant in such an argument.

Reliability

The meanings and treatment of the term 'reliability' exacerbate the situation. 'Reliability' in psychometrics means consistency. Given this meaning, a bathroom scale that consistently reads ten pounds low is a thoroughly reliable scale. A test that asks students to spell ten words, if it gives consistent results, is a psychometrically reliable test, even if the test is called a reading test—or a critical-thinking test for that matter. For psychometric reliability, it does not matter what the test is supposed to be assessing. All that matters is that it do whatever it does consistently.

Typical intuitive types of psychometric reliability are inter-rater reliability (consistency among graders of open-ended tests) and test-retest reliability (consistency from one administration to the next). A third, somewhat less intuitive type, is split-half reliability, in which a test is split into two supposedly similar halves, and the two halves are correlated. In this third type, the consistency is the extent to which

the two halves correlate, adjusted for test length. These consistencies do not indicate that the test is assessing what we hope it is assessing. They only indicate the extent to which the test is consistently doing whatever it is doing.

However, in its ordinary non-psychometric sense, 'reliability' represents "a concept much closer to the measurement concept of *validity*," to use the language of Leonard Feldt and Robert Brennan, authors of the chapter "Reliability" (1989: 106), in the third edition of the authoritative *Educational Measurement*, edited by Robert Linn. In other words, 'reliability' in its ordinary sense when applied to tests expresses concern with successfully testing for what it is supposed to test for, that is, test validity, not just with consistency (see Ennis 2000).

A severe problem here is that the public is not generally aware of the meaning shift from the ordinary to the psychometric sense of 'reliability.' Even people who have taken courses in testing forget it. For many people, the tendency is to interpret 'psychometric test reliability' as 'test validity'; I have seen highly placed educational professionals do this. Hence such people are less likely to worry when there is no explicit argument for the situational validity of a test they are considering, if there is psychometric-reliability information.

One passage in the report reads to me as if the Commission is itself vulnerable to making this mistake because it uses the everyday concept of *reliability* in a report that emphasizes testing:

Compounding all of these difficulties is a lack of clear *reliable* information about the cost and quality of post secondary institutions. (USDOE 2006: x, italics added)

By "reliable information" the Commission here does not mean merely consistent information; it means correct information—on which we can depend—that is, valid information (in the ordinary [dictionary] sense of 'valid'). All of this suggests that the Commission, because it deliberately uses 'reliable' in its ordinary sense, might be open to interpreting psychometric reliability information as validity information, especially if test specialists decline to talk about test validity, as the psychometric leadership now suggests.⁷

My advice so far is 1) to make sure that everyone involved understands that psychometric test reliability is not test validity, even though the reliability information is likely to be emphasized because it is easier to gather and looks better on the face of it; and 2) to demand an argument for the claimed extent of situational validity of any test under consideration, and not be satisfied with only (psychometric) reliability information.

Primarily for multiple-choice tests of critical thinking, the reliability situation is made worse by the fact that some other indices of psychometric reliability that are commonly used, at least partly for the

sake of convenience, are measures of the *internal consistency* of a test (e.g., Kuder-Richardson formulas and Cronbach's alpha). Internal consistency, roughly speaking, is the extent to which each item correlates with every other item. To a test maker and appraiser, these indices have the advantage of being easy to obtain—with only one administration of a test and without any thought about test content—in contrast to the greater difficulty of obtaining the test-retest and split-half reliability estimates that also can be used with multiple-choice tests. To elaborate: The two administrations on the same population that would be required for test-retest reliability are difficult to arrange, and on the second administration there are likely to be some dropouts (who need to be handled somehow), and possibly some recollection by the test takers of items from the first administration, depending on the test and the time lapse. Internal consistency psychometric reliability avoids these operational problems. For split-half reliability, someone has to read the test and identify comparable halves. Furthermore, the halves are almost always different anyway, resulting in more trouble for the test maker or appraiser.

Another advantage of the internal-consistency indices for the test-selling test-maker is that there is a fairly simple statistical way to engineer high internal consistency: Go through the test and replace the items that do not correlate well with the total score with items that do, a process that tends to make the test unidimensional. Of course, it also increases the reported reliability (internal consistency), which many people will interpret as test validity, making the test more attractive to much of the unsuspecting test-consuming public.

I realize that the intuitive kinds of psychometric reliability are at least roughly necessary conditions for situational validity of some tests, but they are not sufficient. Furthermore, for tests of multidimensional traits, the internal-consistency kinds of reliability are not even necessary. It depends on how closely the dimensions are related to each other. If the dimensions correlate very highly with each other, then the internal-consistency indices could be useful. But if the dimensions' correlations with each other are low, then the internal consistency indices of reliability would be misleadingly low, as I have elaborated elsewhere (Ennis 2000).

On the basis of my experience teaching critical thinking I think that critical thinking is multidimensional. A student might excel in one aspect of critical thinking, but not another, resulting in a tendency toward lower internal consistency of a critical-thinking test.

I know of only one piece of empirical research that is relevant to the multidimensional issue in critical thinking. Mines (1980) found that a set of internal-consistency index relationships—using Kuder-Richardson Formula #18—for the separate parts of Cornell Level Z (Ennis

and Millman 2005) were generally nearly as high as the index for the whole test. But because the part scores have many fewer items than the total test, the part score indices should have been considerably lower than that of the total score, if the test is a unidimensional test.⁸ Thus these data suggest that this critical-thinking test is a multidimensional test, suggesting in turn that critical thinking is multidimensional—to the extent that the test was situationally valid. These data are presented in tabular form and discussed in Ennis, Millman, and Tomko 2005: 17–18, 40.

Much more research is needed about empirical critical-thinking multidimensionality. For the present, I urge caution in the use of the internal-consistency indices at least unless and until such research is done. To initiate the research process, I suggest that the relationships among at least these three conceptually identifiable dimensions be studied: ability to judge the credibility of sources and observations, ability to judge whether a line of reasoning is deductively valid, and ability to judge whether a best-explanation argument yields a proof beyond a reasonable doubt (or probable proof, or less). There are others, and there are different ways of conceptualizing critical-thinking dimensions, but those three might be a useful starting point.

A content-oriented way that test makers could respond to this test-maker problem is to reduce the multidimensionality of a test by reducing it to a primarily one-dimensional test. Exclusive or almost exclusive emphasis on deductive validity (using 'validity' as it is used in philosophy, mathematics, and logic—as contrasted with psychometric test validity) as the dimension is an obvious choice in this direction because deductive items with keyed answers that are immune to objection are easier to write and grade than other challenging critical-thinking items or prompts. So there is a danger that deductive validity will be very heavily or exclusively used in a critical-thinking test, sacrificing test validity for the sake of internal consistency.

Hertzka and Guilford (1955) published just such a test, which is now out of print. It consisted only of deductive logic items, but the promotional material claimed that it tested for what is commonly known as critical thinking.

So I supplement the previous two warnings with two more. The first two were 1) to make sure that everyone understands that psychometric reliability is not test validity, and 2) to make sure that the argument for the situational validity of a test under consideration is available and that we see it. The third warning is to be aware that internal-consistency reliability might not be particularly high in a good multidimensional test. The fourth is that in response to the problem that generates the third warning, testing only for prowess in judging deductive validity is

an unfortunate, possible result. Judging deductive validity is important, but there is much more to critical thinking than that.

Teaching to the Test

A standard problem with the statewide accountability testing movement in elementary and secondary schools in the United States is that teachers are pressured to, and often do, teach to a test, neglecting many of the important features in their areas of teaching. However, teaching to a test is not undesirable when the test is still situationally valid even if teaching to it occurs. An example is in realistic performance testing, such as testing pilots to see whether they can land an airplane in a cross wind. So far as I know, the best way to teach pilots this skill is to do what amounts to teaching to the test: teach them how to land an airplane in a cross wind. This example makes the point that teaching to a test is not always bad.

Put simply, the issue is whether the test is still situationally valid, even if the teachers teach to it. If so, there is no problem, if not, then there is a problem.

If the test assesses the total subject, including its application if appropriate (as it is with critical thinking) then the danger of undesirable teaching to the test is considerably reduced, but the process might be expensive. If a test assesses only a sample of, or a selection of, the total subject, then either teachers must restrain themselves, or be restrained, from focusing only on those features that they know are going to be assessed or used, or they must not know which features will be assessed.

It is similar with the format. If a particular test format is going to be used, then teachers must restrain themselves, or be restrained, from focusing only on that format, unless the format is ideally suited to critical-thinking assessment, or they must not know which format will be used.

There are too many factors and possibilities to cover in advance. In broad outline, the important thing here is for critical-thinking faculty and administrators to recognize the problem and to make sure that invalidating a test that otherwise would be situationally valid does not occur, that is, to make sure that the test used is not one the situational validity of which will be compromised by teaching to it, or else that teaching to it does not occur. I realize that this is easier said than done. Vigilance is needed.

Value Added, Pre-Post Comparisons, Student Learning Outcomes:
Other Possible Explanations

Accountability, as I understand it in this context, calls for holding an

institution responsible for producing results commensurate with its costs and claims (explicit or implied). The value added mentioned in the report is one type of result and is generally viewed roughly as the difference between the mean (average) critical-thinking scores of the students at the beginning of their college careers and at the end, with the difference given perspective by the degree of variation (standard deviation) in test scores. More about that later. The quantity, value added, of course assumes that the test is a valid measure of critical-thinking prowess in the situation.

Value added is similar to the well-known pre-test/post-test comparisons, which, however, are not adjusted for the degree of variation. On first glance, both seem to be reasonable ways to appraise an institution's contribution to a student's critical-thinking prowess, and to compare institutions, neglecting costs. But there are dangers. One problem is that, without a control group, there are other possible explanations for whatever changes are found, such as maturation, learning from the test on its first administration, and learning critical thinking from life situations apart from the college experience, such as media blitzes, and work or job experience.

Pre/post comparisons are derisively labeled "pre-experimental" by Campbell and Stanley (1963) in their classic treatise on experimental design. I am not as suspicious of pre/post designs as they are, but to give credit to an institution for improvements one must be reasonably able to rule out plausible alternative explanations.

This same problem exists for the concept of "student success outcomes" (USDOE 2006: 4), or "student learning outcomes" (USDOE 2006: 24), which are popular these days, and not distinguished in the report. I shall use the latter wording. To label a test result a "student learning outcome" is to attribute responsibility for the result to the institution seeking to achieve that result. Although in many cases this attribution might well be justified, the attribution must be defended against alternative explanations, especially in the case of critical thinking. Control groups would help, though it would be difficult to identify them and to secure their cooperation.

Because of these other-possible-explanation problems with value added, pre-post comparisons, and student learning outcomes, I would like to add the qualification 'alleged' before 'value added' and 'outcomes' when I mention them in this commentary. However, I would also like to avoid saying that word each time, so henceforth please understand that qualification to be implicitly there. A similar reservation should be expressed when responsibility for pre-post differences is automatically attributed—without a control group.

Another problem is securing comparable and representative groups for the two test administrations (for value added and pre-post com-

parisons). For example, in cross-sectional comparisons (that is, those made in the same academic year), if the freshman group exhibits less critical-thinking prowess than the senior group, the comparison would be biased in favor of the institution if the difference occurs at least partly as a result of dropouts from the senior class during their stay in college. In longitudinal comparisons (same students over a four-year period), the same students must be tested each time (often difficult to do, made much more difficult by dropouts, again biasing the results in favor of the institution). On the other hand, senior malaise might set in and the seniors might underperform, biasing the results against the institution.

Opportunistic selection of people to actually take the test would be a temptation for college officials in order to get better results. Such manipulation happens all the time in the current accountability-testing movement in grades K-12, according to my sources.

If the test does not count toward grades, either or both freshman and senior groups might underperform. Not counting toward grades is almost inevitable if matrix sampling is used. Matrix sampling gives different parts of a whole test to different students. This is done in order to have a more comprehensive test without taking too much of a given student's time, and thus, it is hoped, a more valid test in the situation. Matrix sampling gives a score to the whole group, but not to any particular student, and might lower motivation, depending on other features of the situation. However, Carol Tucker of Educational Testing Service (personal communication) reports that students she has tested often try hard in critical-thinking tests even when there is no impact on grades because the critical-thinking items are novel and interesting. So here is a factor that might or might not provide an alternative explanation of the results. However, matrix sampling generally does lower the stakes for individual teachers and departments, and thus lowers the temptation to manipulate the results.

Transparency and Comparability: Pressures for Manipulation of the Data

The Commission's joint requirements, transparency and comparability, put strong pressures on institutions to look good, thus inviting manipulation by interested parties. Just imagine your undergraduate institution being publicly compared numerically with a traditional rival (or any other institution for that matter) for critical-thinking value added—so that, for example, parents and students can see whether this year MIT will yet once again beat Harvard and Yale in critical-thinking value added. Which university in the Big Ten this year is going to show the most critical-thinking value added? The Department of Education's transparent comparable database that Secretary Spellings and Chair

Miller want to create will contain the supposedly comparable numbers that will enable these comparisons (with privacy protections for students' names and social security numbers), (Arenson 2006: 1A; Chaker 2006: A12; USDOE 2006: 21–22). We can just look it up in *U.S. News & World Report*, which no doubt will summarize it all each year.

The resulting temptation to manipulate the data could involve the basic aspects of higher education, including selection of students for admission, selection of students for testing, instruction (including teaching to the tests), curriculum, course requirements, *etc.*, and might well extend to the conception of critical thinking on which the test or tests are based and to the specific test or tests used. The controversies about specifying and assessing proposed learning outcomes that many of us have already experienced in our local institutions would increase, requiring vigilance.

Incidentally, transparent comparisons of student learning, in the Commission's view, are not limited to institutions; they are also recommended for comparisons of states, so that "state policymakers can make valid interstate comparisons of student learning and identify shortcomings as well as best practices" (USDOE 2006: 24). As a result, I would be wondering each year whether the State of Illinois will again exceed neighboring Iowa in critical-thinking value added. Thus there would be even more pressure for manipulation and adjustments.

Standard Deviation: A Temptation for Subtle Manipulation

The standard deviation is a measure of the variation in scores of a specific group. A currently popular way of indicating the practical significance of a difference between groups (or, within one group, the difference from one test administration to the next) is to compute the ratio of the difference in mean (that is, average) scores to the standard deviation. This ratio is often labeled Cohen's *d*. I believe that the Commission intends "value added" to be measured by Cohen's *d*, but if not, vigilance is still needed in appraising the use of Cohen's *d* because of its current popularity in appraising student progress in critical thinking.

Cohen's *d* may be contrasted with, and in many cases is for practical purposes an improvement upon, statistical significance, which, with a large enough sample, is often obtainable for results that are not practically significant.

Here is a simplified example of how the use of Cohen's *d* can be problematic: The seven scores, 2, 4, 6, 8, 10, 12, and 14 (range: 2 to 14) from College A vary more (have more dispersion) than the seven scores, 5, 6, 7, 8, 9, 10, and 11 (range: 5 to 11) from College B, though both sets of scores have the same mean (average) score, 8. The College A scores have a higher standard deviation (4.0) than those of College

B (2.0). I will not here present the calculations. Suppose further that each person in each college improves by three points from freshman to senior year, so each group's mean score increases by three points to 11. Thus the ratios of mean improvement to standard deviation, which are the value added, are 3 over 4.0 (3/4) or 0.75, and 3 over 2.0 (3/2) or 1.50; College B students have an improvement, according to this way of figuring it, that is twice as many standard deviation units as those of College A (1.50 is twice 0.75, thus double the value added), even though the actual improvements in mean scores are the same. So the amounts of value added in this hypothetical case differ radically (after adjustment of the improvement by division by the standard deviation), even though the actual mean improvements are the same.

To put these numbers in perspective: As David Hitchcock (2004: 197) interprets the results of Pascarella and Terenzini (2005), in studies that have been done, the average value added in critical thinking, as students progress from freshmen to seniors is .64 of a standard deviation (or .64 standard deviations).

One danger is that college admissions officers, or the people who select the test takers, will be tempted somehow to reduce the standard deviation (which is the divisor in the final ratio), thus increasing the value added. Ways to do this include reducing diversity of a tested group, or reducing the diversity of the whole student body. Different institutions might then have the same means and same improvements, but if the standard deviations differ and no compensating adjustments or explanations are made, this ratio-to-standard-deviations method (Cohen's *d*) can give misleading results.

I am not here rejecting Cohen's *d*. On the contrary, I like it. I am just warning about its possible exploitation. Vigilance is needed.

Comparability: Pressure to Have One or Only a Few Standardized Critical-Thinking Tests

Chair Miller disavowed requiring one particular test or small number of tests (Arenson 2006: 1A; Chaker 2006: A12), saying, "There is no way you can mandate a single set of tests" (Arenson 2006: 1A). But one clear additional consequence of a national data base containing transparent comparisons among institutions for the amount of value added is pressure for all institutions to give the same test. How else can "parents and students have . . . [the] solid evidence *comparable* across institutions of how much students learn in colleges or whether they learn more in one college than another" that is recommended by the report (USDOE 2006: 14, italics added)? With only one test, the comparisons will be cheaper, simpler, and clear. If different tests are used, then, although statisticians will construct a conversion system, allegedly converting all tests to the same scale, possibly based at least

on means and standard deviations, the tests will still be different tests and comparisons among colleges using different tests will probably be misleading. For example, we can attempt to construct a conversion scale for the SAT and the ACT, a pair of college admissions tests, or the Miller Analogies and GRE, a pair of graduate admission tests, by converting all scores to have the same arbitrary mean score (such as 50) and standard deviation (such as 10). But if the two tests in each pair are different, as I think they are (and as have all the people I know who have taken both tests of a pair), how are the scores really comparable? The burden of proof is on the converters. I know that assumptions can be made that would supposedly justify such comparisons among critical-thinking tests, but will the assumptions be justified? I am dubious. We must monitor the assumptions.

Thus the ideal of numerical comparability among institutions is suspect. Comparability provides pressure to have one or only a few tests, in order that real comparability can be achieved. But having only one national test would generate insuperable political and economic problems (as the Commission itself apparently sees and as many members of AILACT urged in our listserv discussion in February 2007), and strong temptations to manipulate the results. Having only a few (perhaps two, three, or four) would still result in full comparability—among institutions using the same tests—and thus still strong temptations because each test would be given to large numbers of institutions.

Abandoning numerical comparability as an ideal in critical-thinking value added would greatly reduce the temptations to manipulate the results, and permit individual institutions to choose among a wide range of tests, under the supervision, I urge, of an accrediting agency. Each test does not have to be totally different from every other test. A large testing organization can make various sets of tests that have members that are similar in a variety of ways. But they should not be the same test, thus destroying general numerical comparability, and vastly reducing the temptations to manipulate the data.

It might be acceptable to have a few exceptions endorsed by the institutions involved for their own reasons. For example, the University of Illinois and the University of Iowa might use the same test, agreeing in advance to resist the strong temptations to manipulate. It might be possible, as in a football match between these two universities, to have very strict rules and referees, assuring the absence of manipulation. However, I am dubious because of the non-public nature of teaching and administration, but will not unequivocally rule out the possibility. Extreme vigilance would be required.

I realize that some people will try to make different tests appear comparable by some arbitrary conversion system, and will claim that they are comparable "for all practical purposes." If so, then the

comparability-induced pressures and temptations for manipulation of the situations and the data will develop and do much damage to higher education. It would be worse than at the elementary and secondary levels of education because these levels for the most part have a captive audience. Higher education is almost totally dependent on attracting students. Much vigilance is required: if unsuccessful, then critical thinking faculty and administrators should reject the Commission's program.

Test Ceilings and Floors: Another Problem with Comparability

If we were to have comparability between institutions, the tests would have to be the same (or truly comparable—unlikely in my opinion, as I have argued). A test that is appropriate for the middle range of students, which is the most likely situation, and produces comparable results for midrange students, will probably have a ceiling or partial ceiling for top students. So a college overloaded with top students might not show much, if any, improvement, even if its students did superbly on the test both times, and even though its students made vast improvements in unassessed advanced linguistic understanding and sophistication, such as dealing with impact equivocation (Ennis 1996: 353–60) and operational definition (Ennis 1969). Similarly, a college with a large number of lower-range students might show little improvement because many of its students have such difficulty with the questions that their answers would approach randomness, even if they had made great progress at levels of critical thinking deemed too easy to have been assessed on a mid-range test and thus below the floor of the test.

There are ways of supposedly putting all levels on one scale, as in computer-adaptive testing, a relatively recent development in which, roughly speaking, the first few items determine the student's rough level and the next items (selected by a computer from a set of items arranged in order of difficulty) successively refine that determination, so that a student theoretically does not even see items that are much too difficult or too easy for her or him.

For multiple-choice tests of multidimensional concepts in which the dimensions do not correlate highly with each other (like critical thinking, as I have suggested), there is no way all items can be ranked in a single list for assignment by a computer because students differ in their prowess on different dimensions. So unidimensional computer-adaptive testing seems inappropriate for a multidimensional concept like critical thinking, if the dimensions are not highly intercorrelated. Multidimensional computer-adaptive testing is theoretically possible, but its adaptability to critical thinking is to my knowledge unexplored

and would be complicated. So I urge monitoring the way this problem is handled.

For open-ended testing, arranging prompts in order of difficulty seems unlikely because prompts vary so much from one to the next and are also very likely to be multidimensional. The more authentic the prompts or requests, the greater the difficulty.

Test Making, Administration, and Scoring

Although it is tempting to do the testing locally, the high-stakes nature of this testing will require extensive security, continual making of new forms of the tests, and very careful development and testing of each form, if the test used is multiple-choice. If the test is open-ended, for example, by means of essays, then there are still problems with security and replacement and testing of the prompts, passages, or topics, even if the scoring rubrics are generic. There are also validity and psychometric reliability problems for open-ended assessment (including authentic performance assessment). There is a danger that scoring will not be consistent across graders, and also that important aspects of critical thinking will not be assessed. For example, essay tests, for a given amount of a student's time, tend to be less comprehensive than multiple-choice tests. They are more likely to neglect specific aspects of critical thinking.

Because of the expense of scoring open-ended tests it is of note that there is now a computer-grading approach to essay testing, called "E-Rater," developed by Educational Testing Service. I am not familiar with the details of its operation, but am dubious about computer scoring of critical-thinking essay tests, based on my own experience in developing and grading critical-thinking essay tests. I do not see how a computer, which can only follow explicit directions, could deal reasonably with the frequent unpredicted responses to the prompts in the tests on which I have worked.

Both of the tests mentioned earlier as recommended by the Spellings Commission (CLA and MAPP) use E-Rater for responses to writing prompts. My advice here is to be vigilant if E-Rater is used for critical-thinking essay testing. For example, results should be checked against local human-expert evaluations for samples of students—on the same prompts (essay assignments)—using an acceptable and publicly available conception of critical thinking.

Whether multiple-choice or open-ended assessment is used, the cost of the assessment in high stakes situations will vary roughly with the quality. Furthermore, the quality needed for high-stakes testing, no matter whether it is multiple-choice, or open-ended, or some combination, will require large organizations with experience in quality assessment of critical thinking and sufficient qualified personnel to do the devel-

opment, to administer the test, to do or supervise the scoring, and to maintain security, though possibly test development can be handled by experienced qualified smaller units with the cooperation of a large organization for tryouts.

To the extent that strict comparability can be relaxed, as I urge, the demands accordingly would be less, but I believe generally would still be too great for individual educational institutions to be the test developers.

Incentives, Accreditation, Costs, Consequences

In its final version, the Spellings Commission report did not call for direct national governmental requirements, enforcement, or threats of withholding of federal funds and grants made for other purposes. Instead, Secretary Spellings mentioned matching funds provided to colleges, universities, and states that collect and publicly report the outcomes (Chaker 2006: A12). The report itself speaks of "incentives for states, higher education associations, university systems and institutions" to cooperate (USDOE 2006: 24). So the recommended approach appears to be a carrot, not a stick, approach, at least for the time being.

In addition to using incentives or matching funds to exert pressure, the report urged accreditation agencies to exert pressure, giving priority to transparent comparable performance outcomes, as opposed to inputs and processes (USDOE 2006: 25). Inputs and processes are such things as course requirements, money spent on teaching, workshops for teachers, *etc.* Performance outcomes include an institution's critical-thinking "value added," as determined by tests (accompanied by the problems I have mentioned already).

The incentives and matching funds would, I presume, leave at least half the expense to the institutions whose students' critical thinking is assessed. The accrediting agencies' expenses would also, I presume, be paid by the institution being reviewed. So the danger is that, even if locally controlled, institutions with tight budgets (which are most of them) will be reluctant to spend the money necessary to do a decent job of assessing students.

Dangers in Not Testing for Critical Thinking; Transparency

Because of problems and dangers I have mentioned and others as well, some critical-thinking specialists will urge that we avoid any cooperation with the report's advocacy of nationwide testing of critical thinking. But there are also dangers in this position:

One danger is that what is not tested is less likely to be taught and learned in an accountability-testing environment. This is a problem

that a number of curriculum areas have felt in the pervasive statewide elementary and secondary testing programs now in existence in the United States.

A second danger is that content masquerading as critical thinking will continue to be taught under the banner of critical thinking, and that its promoters will claim to be pursuing the critical-thinking mission of their institutions. In some cases, this content might be justified as higher education content. But to adopt it under a misnomer misleads the public into mistakenly thinking it is getting something it is not getting.

If there is controversy about whether the assessed content in a test is critical-thinking content, let the informed public judge—with full transparency of the competing conceptions, the tests used to assess them, and the validity arguments—and then vote by their actions.

Better enabling wise judgments by higher-education clients is one reason for full transparency. A second reason is that full transparency better enables critical-thinking faculty and administrators, as well as policy makers, to make wise decisions. If we do not know the details about the prospective tests, how can we choose among them, and how can we adequately guard against manipulation of results?

A Stance Toward the Critical-Thinking Aspects of the Report

Partly because what is not assessed is less likely to be taught, partly because epistemic critical thinking is so important, and partly because I believe that there are a number of courses masquerading as critical-thinking courses that are not critical-thinking courses; I recommend, in spite of the many reservations I have indicated, that critical-thinking faculty and administrators for the most part go along with the recommendations of the report in the area of critical thinking.

One crucial reservation accompanying this recommendation is that I do not endorse the strong comparability pressures that seem to call for only one test or a few tests (although Chair Miller denies seeking only one test or a few tests). Instead I would urge a large number of different noncomparable tests, the choice among the available ones lying with each institution (monitored by the accrediting agency), resulting in medium stakes (rather than high stakes) testing, in lower costs, and in much less temptation to manipulate results.

If anyone tries to make different tests appear comparable by using an arbitrary conversion system, this effort should be strongly resisted, not only because the tests would not really be comparable, but the alleged comparability would invite manipulation of the data and the process in order to make an institution look good when compared with others.

However, the choice in each institution should be institution-wide, as AILACT members have urged, rather than made by different units or individual teachers within an institution. This institution-wide decision would reduce the danger of promotion of things under the label 'critical thinking,' that are not critical thinking in the standard sense, and would help to maintain a standard of quality of teaching. One danger then would be that institutions might choose the test to make them look good. So the situation requires full transparency, enabling vigilant critical-thinking faculty and administrators to discern and challenge difficulties.

There should be much more transparency than the report specifies—to include transparency of conceptions of critical thinking, of the argument for the claimed extent of the situational validity of the tests being considered, and of the tests themselves, generally by means of no-longer-used (but still representative) versions of the tests, so that critical-thinking faculty, administrators, and clients (policymakers, parents, students, etc.) can make informed choices with full transparency.

Accepting my recommendations requires much vigilance on the part of these people, especially critical-thinking faculty and administrators.

Summary and Comment

The Spellings Commission report of 2006 advocates many changes in higher education but I have focused on one crucial area of the recommendations: transparent, comparison-enabling value-added accountability testing of critical thinking that would supposedly enable parents, students, and policymakers to make comparisons of, and informed choices among, colleges and universities, and intelligently decide whether they are getting their money's worth. In addition to critical thinking, literacy and perhaps mathematics are also emphasized, but I have not focused on them, though the situations are similar.

My concerns in this essay are not limited to the Spellings Commission report or to the United States. These concerns exist for testing all over the world, including testing required or urged by states, provinces, and other governing bodies, and by accrediting agencies.

A number of problems and dangers could arise in the implementation of the Commission's recommendations for critical thinking, including these:

- a) neglect of the everyday, epistemic concept of critical thinking;
- b) neglect of the distinction between psychometric reliability and situational test validity;
- c) overemphasis on psychometric reliability at the expense of situational test validity;

d) the possibility that internal-consistency methods of determining psychometric reliability will discriminate against multidimensional critical-thinking tests;

e) more specifically, the possibility that in order to keep up the psychometric reliabilities, a critical-thinking test will be limited to deductive logic;

f) neglect of other possible explanations of value-added evidence, pre-test/posttest differences, and "student-learning-outcomes" evidence;

g) failure to have fairly large organizations for administering, scoring, and providing test security—with the concomitant need for qualified personnel for making and supervising the use of a critical-thinking test;

h) choosing (or on the part of higher education's clients, respecting) a test simply because it is labeled "critical-thinking test" (more deeply, making a decision about a test and its relevance without having full transparency, which would include making available the nature of a test and the conception on which it is based, sample tests, and the argument for the situational validity of a prospective test);

i) sacrifice of test quality for the sake of economy;

j) the use of computers to do the grading of essays solicited in order to assess critical thinking, and the possible failure by local people to check the results of such testing;

k) the wrong kind of teaching to the test, the kind that destroys a test's situational validity;

l) the use of standard deviations to exhibit practical differences in test results, inviting manipulation;

m) the difficulty of actually achieving comparability among institutions using different tests, resulting in pressure for a single, or a few national tests;

n) the difficulty of achieving comparability, even if only one test were used, because of floors and ceilings of tests.

o) the severe political and economic problems of a single required test, including temptations to manipulate results; and

p) manipulation temptations arising from the prospect of numerical comparisons among institutions using supposedly comparable tests.

Actually the last six (k through p) dangers or problems result from or are exacerbated by the goal, comparability, which I urge critical-thinking faculty, administrators, and clients to reject.

In spite of these problems and dangers, Commission-inspired critical-thinking testing might well help produce more attention to critical thinking in individual courses and in the total curriculum, a check on whether critical thinking in the everyday sense (the epistemic sense) is being taught, a check on the effectiveness of critical-thinking in-

struction, and helpful information for parents, prospective students, and policy makers.

I recommend that critical-thinking faculty and administrators cooperate with the Commission, but support much less comparability among institutions than the Commission proposes, and seek much deeper transparency than recommended by the Commission. The transparency should include details about the concept of critical thinking on which a test is based, samples of the test, testing procedures, and the argument for its situational validity.

Furthermore, I urge proactive vigilance by critical-thinking faculty and administrators toward the many problems and dangers I have broadly indicated, and which will vary in their specific manifestations from situation to situation. If the vigilance is not successful, then I would not recommend cooperation.

There is much more to say on both sides about all of these topics. I have only barely touched the surface for some, and trust that there will be a continuing dialogue. I hope to have helped open the doors for a reasonable and vigilant response by critical-thinking faculty, administrators, and clients to this rather radical proposal by the Spellings Commission, a proposal that recognizes the fundamental importance of critical thinking.

Notes

I deeply appreciate the extensive helpful comments about earlier versions of this essay by Carol Tucker, Jennie Berg, Michael Scriven, Donna Engelmann, Sean Ennis, and Fred Ellett. One earlier version was presented at the Central Division Meeting of the American Philosophical Association, Chicago, April 19, 2007.

1. Named for Margaret Spellings, U.S. Secretary of Education. The report is entitled "A Test of Leadership: Charting the Future of U.S. Higher Education" (U.S. Department of Education 2006).

2. Because the Commission seeks comparability in its database (USDOE 2006: 14, 21–22), I believe it means 'numerical comparability,' which is the kind against which I argue. In order to save space and avoid repetition, I shall often omit the qualifier, 'numerical.'

3. In both likely interpretations of "critical-thinking faculty and administrators": those who think critically, and those who also teach or otherwise promote critical thinking.

4. Again, in both likely interpretations of "critical-thinking faculty and administrators."

5. I realize that I have not in this paper given an extensive argument for the conclusion that the everyday notion of critical thinking is the epistemic one. For the most part, I here simply assume it, but do invite readers to consider seriously this point about what employers might have meant in complaining about the lack of critical thinking among their employees.

6. They urge that validity is not a property of tests at all, but rather a property of inferences drawn from, or interpretations of, test scores (Messick 1989a and 1989b; Joint Committee 1999; Kane 2006). I think this is confusing, and deprives us of a special phrase ('test validity') to express our overarching concern, as teachers, about the extent

to which an educational test in a given situation successfully assesses what we want it to assess. Admittedly, one particular inference from, or interpretation of, test scores is about the extent to which test scores are successful indications of whatever it is that the test is supposed to assess in a situation, which roughly amounts to the extent to which the test is situationally valid. But the range of possible inferences from, or interpretations of, a set of test scores is very broad. Hence, in this new theory of validity in testing, our concern about whether a test in a given situation assesses what it is supposed to assess is no longer deeply, conceptually established as an overarching concern. My extended argument regarding this difference appears in a paper in which I provide a broader contextual definition of 'situational validity' than I do here—to cover more than educational tests (in process).

7. See previous note.

8. By the Spearman-Brown formula, if a homogeneous test has fifty-two items and an internal-consistency reliability of 0.76 (which are the numbers for this test in that administration), a similar test of only four items should have an internal-consistency reliability of 0.16. But the three parts of this test that each had four items actually had internal-consistency reliabilities of 0.72, 0.65, and 0.60. The ten-item part had an obtained index of 0.76 (the same as the total test), though the predicted index for a ten-item similar test, assuming homogeneity, is 0.35. All of these numbers are only estimates, and there is no replication, but the differences from the predictions that assume homogeneity and thus unidimensionality are substantial.

Bibliography

- Arenson, Karen W. 2006. "Panel Considers Standard Tests for Colleges," *Sarasota Herald Tribune* (February 9): 1A.
- Campbell, Donald T., and Julian C. Stanley. 1963. "Experimental and Quasi-Experimental Designs for Research on Teaching," in *Handbook of Research on Teaching*, ed. Nathan L. Gage. Chicago: Rand McNally, 171–246.
- Chaker, Anne Maric. 2006. "Spellings Aims to Impose New Standards on Colleges," *Wall Street Journal* (September 27): A12.
- Council for Aid to Education. N.D. *Collegiate Learning Assessment: CLA in Context*. http://www.cae.org/content/pro_collegiate.htm, accessed April 13, 2007.
- Educational Testing Service. 2007. *MAPP: The Measure of Academic Proficiency and Progress, User's Guide Draft*. Princeton, N.J.: Educational Testing Service.
- Ennis, Robert H. 1962. "A Concept of Critical Thinking," *Harvard Educational Review* 32: 81–111.
- _____. 1969. "Operationism Can and Should Be Divorced from Covering-Law Assumptions," in *The Nature and Scope of Social Science: A Critical Anthology*, ed. Leonard J. Krimerman. New York: Appleton-Century-Crofts, 431–44.
- _____. 1987. "A Taxonomy of Critical-Thinking Dispositions and Abilities," in *Teaching Thinking Skills: Theory and Practice*, ed. Joan Baron and Robert Sternberg. New York: W. H. Freeman, 9–26.
- _____. 1989. "Critical Thinking and Subject Specificity: Clarification and Needed Research," *Educational Researcher* 18:3: 4–10.
- _____. 1990. "The Extent to which Critical Thinking Is Subject Specific: Further Clarification," *Educational Researcher* 19:4: 13–16.
- _____. 1991. "Critical Thinking: A Streamlined Conception," *Teaching Philosophy*, 14:1: 5–25.
- _____. 1992. "John McPeck's Teaching Critical Thinking," *Educational Studies* 23:4: 462–72.
- _____. 1996. *Critical Thinking*. Upper Saddle River, N.J.: Prentice Hall.
- _____. 2000. "Test Reliability: A Practical Exemplification of Ordinary Language Philosophy," in *Philosophy of Education 1999*, ed. Randall Curren. Urbana, Ill.: The Philosophy of Education Society, 242–48.
- _____. 2002a. "Goals for a Critical-Thinking Curriculum and its Assessment," in *Developing Minds*, ed. Arthur L. Costa, 3rd Edition. Alexandria, Va.: ASCD, 44–46.
- _____. 2002b. "An Outline of Goals for a Critical-Thinking Curriculum and its Assessment," <http://faculty.ed.uiuc.edu/rhennis/outlinegoalsctcurassess3.html>.
- _____. 2004. "Applying Soundness Standards to Qualified Reasoning," *Informal Logic* 24:1: 23–39.
- _____. In press. "Investigating and Assessing Multiple-Choice Critical-Thinking Tests," in *Critical Thinking, Education and Assessment*, ed. Jan Sobocan and Leo Groarke. London, Ont.: Althouse Press.
- _____. In process. "Situational Test Validity."
- Ennis, Robert H., and Jason Millman. 2005. *Cornell Critical Thinking Test: Level Z*, 5th ed. Seaside, Calif.: The Critical Thinking Co.
- Ennis, Robert H., Jason Millman, and Thomas N. Tomko. 2005. *Cornell Critical Thinking Tests: Administration Manual*, 5th ed. Seaside, Calif.: The Critical Thinking Co.
- Facione, Peter A. 1990. "The Delphi Report," executive summary of "Critical Thinking: A Statement of Expert Consensus for Purposes of Educational Assessment and Instruction." Milbrae, Calif.: California Academic Press. Available at http://www.insightassessment.com/pdf_files/DEXadobe.PDF.
- Feldt, Leonard, and Robert Brennan. 1989. "Reliability," in Linn 1989: 105–46.
- Fisher, Alec, and Michael Scriven. 1997. *Critical Thinking: Its Definition and Assessment*. Point Reyes, Calif.: Edgepress.
- Harman, Gilbert H. 1973. *Thought*. Princeton, N.J.: Princeton University Press.
- Hertzka, Alfred, and J. P. Guilford. 1955. *Logical Reasoning*. Orange, Calif.: Sheridan Psychological Services.
- Hitchcock, David. 2004. "The Effectiveness of Computer-Assisted Instruction in Critical Thinking," *Informal Logic* 24:3 (Fall): 183–217.
- Johnson, Ralph. 1996. *The Rise of Informal Logic*. Newport News, Va.: Vale Press.
- Joint Committee on Standards for Educational and Psychological Testing of American Educational Research Association, American Psychological Association, and National Council on Educational Measurement (Joint Committee). 1999. *Standards for Educational and Psychological Tests*. Washington, D.C.: American Educational Research Association.
- Kane, Michael T. 2006. "Validation," in *Educational Measurement*, 4th ed., ed. Robert L. Brennan. Westport, Conn.: American Council on Education and Praeger Publishers, 117–64.
- Linn, Robert L. 1982. "Ability Testing: Individual Differences, Prediction and Differential Prediction," in *Ability Testing: Uses, Consequences, and Controversies*, ed. A. K. Wigdor and W. R. Garner, Part 2: Documentation Section. Washington D.C.: National Academy Press, 335–88.
- _____. ed. 1989. *Educational Measurement*, 3rd edition. New York: American Council on Education and Macmillan.
- McPeck, John E. 1981. *Critical Thinking and Education*. New York: St. Martin's Press.
- Messick, Samuel. 1989a. "Meaning and Values in Test Validation: The Science and Ethics of Assessment," *Educational Researcher* 18:2: 5–11.

- _____. 1989b. "Validity," in Linn 1989: 13-103.
- Mines, R. A. 1980. "Levels of Intellectual Development and Associated Critical Thinking Skills in Young Adults," *Dissertation Abstracts International* 41: 1495A.
- National Center for Educational Statistics (NCES), U.S. Department of Education. 1995. *National Assessment of College Student Learning: Identifying College Graduates' Essential Skills in Writing, Speech and Listening, and Critical Thinking*. Washington, D.C.: U.S. Government Printing Office.
- Pascarella, Ernest T., and Patrick Terenzini. 2005. *How College Affects Students: Findings and Insights from Twenty Years of Research*. San Francisco: Jossey Bass.
- Paul, Richard. 1993. *Critical Thinking: What Every Person Needs to Know to Survive in a Rapidly Changing World*. Santa Rosa, Calif.: Foundation for Critical Thinking.
- Siegel, Harvey. 1988. *Educating Reason: Rationality, Critical Thinking, and Education*. New York: Routledge, Chapman & Hall.
- U.S. Department of Education (USDOE), 2006. *A Test of Leadership: Charting the Future of U.S. Higher Education*. Washington, D. C.: U.S. Government Printing Office.
- Robert H. Ennis, 3904 Trentwood Place, Sarasota FL 34243; rhennis@uiuc.edu