A Field Guide to Critical-Thinking Assessment

KEVIN POSSIN
Winona State University

Abstract: A non-technical guide to some of the popular methods and tests for assessing how well students are acquiring critical thinking skills in their courses, programs, or college careers.

The American Philosophical Association's "Statement on Outcomes Assessment" (OA) (APA 2001) nicely captures the trickle-down impetus for doing outcomes assessment:

> OA on the institutional level attempts to determine how well an institution prepares its students to enter the endeavors for which it is presumed to seek to train them. Such matters as accreditation and the availability to its students of federally guaranteed loans may rest upon the demonstrated success of a college in doing so. Because the success of an institution in these matters is of considerable practical importance, and because this success in turn depends to a great extent upon the effectiveness of individual programs, many college administrators have begun requiring departments to develop comprehensive plans for implementing OA. So, for example, departments may be asked to demonstrate "objectively" the differences their degree programs make in the development of students' abilities. In turn, instructors may be asked to formulate specific goals for each of their courses, and to develop instruments to measure their attainment.

And when it comes to actually *doing* OA, the APA takes a rather skeptical position, which can be summarized as follows: What exactly do we need to assess outcomes for? That's the function of those exams and writing assignments we use to determine course grades. If we didn't think they accurately measured how much the students had learned in the class, we wouldn't have assigned them in the first place!

An administrator at my university once expressed his exasperation with such "faculty who think that exams actually measure learned outcomes and value added." Strange, isn't it? Behind this appears to be the belief that exams and assignments, as they are currently used by faculty for determining grades, are irrelevant and that some sort of outcomes

assessment instruments, *not* to be used for grading purposes, are the ultimate representations of student learning. The cynic in me wonders, however: if faculty were to take their usual means of evaluation and take some external set of "OA instruments" and *switch* them, would the irrelevant magically become relevant and the relevant irrelevant in the eyes of those calling for accountability through OA?

It was for the sake of "accountability" that the recently formed Spellings Commission on the Future of Higher Education optimistically entertained the possibility of requiring a national standardized OA test designed specifically for college students, to verify their writing, critical-thinking, and problem-solving skills. No college student left behind, so to speak.

In response to the Commission's report (Spellings Commission 2006), however, Robert Ennis drafted a Resolution, passed by the Association for Informal Logic and Critical Thinking (AILACT) overwhelmingly recommending "strong opposition to having one nationwide standardized critical-thinking test" and recommended "that its members and other endorsers of critical thinking be vigilant in monitoring the extent to which various problems and dangers occur in testing for the 'value added' by higher education institutions to students' critical thinking, and take appropriate action when they do occur" (AILACT 2007). Ennis's detailed discussion of these "problems and dangers" recently appeared in this journal (Ennis 2008).

I've talked with many faculty who think that the *only* test that could measure one's acquisition of cognitive abilities is *life itself*—no less an open-ended, real-world, problem-solving project will do. (I suppose getting mentioned in *News of the Weird* or receiving a *Darwin Award* would qualify as a low score.) Trudy Govier (1987), for example, questions the core motives behind using OA tests to measure *critical thinking* abilities specifically:

> The supposed need for [CT assessment] tests comes from interests of bureaucratic efficiency and academic-political lobbying, not from truly educational, philosophical, or critical interests. . . . [W]e are caught in a trade-off situation with these tests. Perhaps the way out of this dilemma is to refuse the task. (Govier 1987: 268)

I'm not quite *this* skeptical, however, since I've discovered a few assessment tests that *are* acceptable in measuring how effective a course or a curriculum, or even an entire four-year college education, is at enhancing the CT skills of students. So this is my goal here, in keeping with Ennis's and AILACT's recommendation: to critically review some of the commercially available and internationally recognized CT assessment "instruments" and make further recommendations—one lowly faculty to another—that I hope will be useful to those of you who are also feeling the urge (or the pressure) to demonstrate "value added" in your courses, programs, or institutions regarding CT skills. *Better we do our own assessment, lest it be done for us.*

First a Little Philosophy

Different CT assessment tools focus on different CT skills (when they focus on CT skills at all). It's very important to match the assessment test you choose with the CT skills studied in your curriculum, to avoid getting an inaccurate measurement of results—like using your gas gauge to measure your engine temperature. This coordination of CT assessment tests and CT curricula is best achieved by first asking the question, "What is critical thinking?"

With this question, there is a real danger that we may *never* even *get* to our project of discussing OA tests. There has been a robust, and not always friendly, debate about what CT is. At one end of the spectrum there are some who think it is basically formal logic, and at the other end are some who think it is merely meta-cognition. Formal logic, however, is too often focused on *only* deductive relations among symbols, with no attention to their useful application to argumentation, belief formation, and problem solving as found in daily life, often in the form of inductive reasoning. And meta-cognition is too easily construed as merely reflecting on what one happens to be thinking or experiencing at the time, with little substantive guidance as to how to do so, other than with, *e.g.*, open-mindedness, fair-mindedness, clarity, precision, depth, breadth, and logicalness (Paul and Elder 2001a) or by means of knowledge acquisition, comprehension, application, analysis, synthesis, and evaluation (Bloom 1956). Using the meta-cognition account, my university all too easily approved a *dance* class as a CT course, because its students self-reflect on their bodily movements.

Some, for example McPeck (1981, 1991) and Weinstein (1995), think that, since CT is always thinking about *something*, it is incommensurably subject-specific, with no *generic* standards applicable across disciplines. But the transferability of CT skills, which most of us believe is a major goal or benefit in acquiring them, indicates how questionable this position is—certainly denying the consequent is valid no matter what the topic of the argument, and an ad hominem is fallacious no matter who is being attacked (*e.g.*, scientist, politician, historian, or philosopher). That CT is always contextual and context sensitive does *not* entail that it is content *specific*. This is recognized by, for example, Paul and Elder, in their definition: "CT is that mode of thinking—about any subject, content, or problem—in which the thinker improves the quality of his or her thinking by skillfully taking charge of the structures inherent in thinking and imposing intellectual standards upon them" (Paul and Elder 2001a: xx). This definition, unfortunately,

still doesn't answer our original question; it merely rephrases it as, "What are the intellectual standards of CT?" McPeck (1981: 7) also states that CT is "the propensity and skill to engage in an activity with reflective skepticism." Besides being self-defeating, however, this too doesn't explain what such a (generic?) skill is.

Some would analyze CT more in terms of the skillful activity of persuasion, conflict resolution, or debate, *e.g.*, van Eemeron, Grootendorst, and Snoek Henkemes (1996) and Walton (1989). However, the degree to which these rhetorical, dialectical, or dialogical accounts of CT focus on the goal of *merely* getting others to adopt one's beliefs, values, and action plans is the degree to which these views part ways with the fundamental goals of CT—pursuing the truth and avoiding error.

Well, it's time I face the question and briefly present what I hope is a very modest and standard account of what critical thinking is. Essentially, CT is *the practice of requiring, assessing, and giving cogent reasons for one's beliefs, values, and actions* (Possin 2002a). CT, then, can roughly be analyzed into the following component skills or competencies:

- Identifying *reasons* or *arguments*.
- Dissecting arguments into *premises, conclusions*, and *subconclusions* (explicit and implicit).
- Taxonomizing arguments as *deductive* or *inductive*.
- Assessing the *cogency* of arguments, *relative to their type*, in terms of the truth or *acceptability* of their premises and the *relevance* of their premises to the truth or probable truth of their conclusions.
- Identifying *formal* and *informal* fallacies—*i.e.*, popular ways of failing these cogency conditions.
- Critically reviewing *definitions* and *analyzing concepts*.
- Assembling these competencies so as to select and argue for positions on a diversity of issues and critically review competing positions and their arguments, all in a cogent and intellectually honest manner.

This account of CT closely mirrors the consensus view expressed in the famous APA-sponsored "Delphi Report" (1990), and its constituent skills match those outlined by Robert Ennis (2002). It captures at least a *core* curriculum of CT: "on virtually any account of critical thinking, deductive competence, linguistic sensitivity, inductive competence, and the ability to detect fallacies would constitute minimally necessary conditions of critical thinking" (Govier 1987: 254).

Some, however, *e.g.*, van Eemeron, Grootendorst, and Snoek Henkemes (1996) and Johnson and Blair (1991), might fault my analysis of CT and its component skills by claiming that it is instead an analysis

of *informal logic* (IL). For example, Johnson and Blair (1987) define IL as "a branch of logic whose task is to develop non-formal standards, criteria, procedures for the analysis, interpretation, evaluation, criticism and construction of argumentation in everyday discourse." But, after reading the following passage from these authors, one can see why many of us, *e.g.*, Ennis (1984), still can't seem to find a *substantive* difference between CT and IL:

> "Informal Logic" denotes a loosely defined *field of inquiry*, centered on developing an adequate theory for the interpretation and assessment of arguments. In our view an undergraduate informal logic course should teach students the current theory and help them improve their skills in application. "Critical thinking" denotes a *moral/intellectual virtue*—the intellectual activity of thinking critically and the moral disposition to engage in it. Critical thinking courses should teach students the theory and skills, and inculcate the attitude, required to exercise this virtue. One candidate for inclusion in the critical thinking syllabus—among others—is the theory and skills set of informal logic. (Johnson and Blair 1991: 50)

Besides providing only a circular description of CT here, Johnson and Blair admit that IL *can* be a subset of a CT curriculum. So what, then, is *unique* to the CT curriculum? Nothing, judging from Claude Gratton's (2001) impressive comparison of over a hundred CT and IL texts.

Moreover, many *candidates* for a distinction between CT and IL are refuted by Jan Sobocan (2003), in a thorough review of the literature. She concludes that the distinction is merely "semantic." For example, CT, unlike IL, is *said* (above) to focus on enhancing moral/intellectual *virtues* or *dispositions*. But the *skills* and *practice* of assessing and constructing good arguments or reasons *are* dispositional, and the resultant practices and goals of pursuing the truth and avoiding error certainly qualify as moral/intellectual virtues.

CT is *said* to instill broad *self*-directed "intellectual virtues," instead of IL's mere argument-assessment skills. But whenever these virtues are unpacked more explicitly, it's always in terms of conditions for cogent reasoning and argumentation applicable to oneself as well as others.

CT, unlike IL, is also *said* to extend to cases of *non*-argumentation, for example, to problem solving (Paul and Elder 2001a) or decision making or aesthetic judgments (Govier 1987) or graphic advertising. But it is quite reasonable to *reconstruct* problem solving and decision making in terms of drawing *conclusions* about what one ought to do given the *relevant information* (premises) regarding one's circumstances and goals. And one can likewise *reconstruct* many aesthetic judgments and graphic advertisements in argumentative form, as *reasons* for believing in the worth of an artwork or a product—as *why* one *ought (not)* to look or listen or purchase. In the case in which an aesthetic judgment might *not* lend itself to being reconstructed as an argument,

it might well be because it is a mere subjective expression of one's likes or dislikes or what Arnold Isenberg (1949) (misleadingly) calls a "critical communication," analogous to pointing at the work and urging others to "look at *this*, in *this* way and *that*!" In neither of these cases, however, is there any critical thinking, involving the application of "intellectual standards."

Well, I *warned* you that addressing this question about the nature of CT was dangerous. I know I haven't satisfied all of you with my account of CT and my list of its constituent aspects or skills. I apologize for that and admit that, as a result, my forthcoming assessment of CT assessment tools *in light of this account* may not strike some of you as entirely useful. For the rest of you, however, let's resume our discussion of matching one's means of assessment to one's curriculum, with the goal of achieving an accurate measurement of acquired CT skills.

Many current CT and introductory/informal logic courses seem to organize their study of CT skills under three general areas:

1. Informal and formal logic
2. Inductive reasoning
3. The construction of cogent argumentative essays

There usually isn't enough time in a single semester, however, to adequately study all three of these areas. With mostly informal, but some formal, logic constituting the *core* CT curriculum, the fork in the road is most often whether to study inductive reasoning or study how to construct (and critically review) argumentative essays on a variety of issues (after all, a position paper is just critical thinking about an issue, written down). This latter route often makes the course writing intensive, perhaps in conjunction with a freshman composition program (*e.g.*, Hatcher and Spencer 1993). Although this is not necessarily the case: for instance, after discussing the essential elements and various formats of a position paper (Possin 2002b), I have my students do "Anatomy of a Position Paper" exercises on numerous op/ed pieces, in which they must identify the parts of a position paper being presented in each paragraph (*e.g.*, statement of position, argument for position, criticism of alternative position).

Before I begin discussing how these coarse- and fine-grained differences in the contents of CT curricula are best matched with available CT assessment tests, let me first clear out some dead wood concerning *other* means of assessment.

Assessment Surveys

The self-reporting survey is a very popular assessment tool, because it's so easy. In fact it's *too* easy. It basically works as follows (to the tune of the *William Tell Overture*, if you like):

Survey: Do you think you've learned a lot?
Student: Yes, I think I've learned a lot.
Faculty: Then you must have learned a lot.
Administrator: Yes, they must have learned a lot!

Surveys are notoriously unreliable indicators of actual competencies. The Association of American Colleges and Universities (2005) found that "standardized test results indicate that only 6 percent of seniors graduate at the 'proficient' level in critical thinking skills, while 87 percent of students believe that college contributed a great deal to improving their skills in this area." (Faculty surveys fare no better: the same study found that 93 percent of the faculty surveyed report making CT a focus of their courses.)

This chasm between students' perception and reality can all too easily happen to whatever degree academic quality and rigor are sacrificed to unchecked concerns about, *e.g.*, self-esteem (Stout 2000) or high retention rates (Dresner 2004). Instructors often feel the pressure to win more favorable student survey ratings either by lowering their academic standards or increasing their assigned grades. I did an informal analysis of the 9,290 ratings on Ratemyprofessors.com for the faculty at my university and found that the strength of correlation between the *perceived quality* of a course and its *perceived ease* was .59, which is quite high. (Admittedly, Ratemyprofs is a self-selected sample; but then so is *every* non-mandatory student-assessment survey.) In a similar study involving twenty-five colleges (3,190 faculty with ten or more postings, for a total of 65,678 postings), this strength of correlation was .61 and steadily increased with the number of postings until it reached .93 for those faculty with ninety or more postings (Felton, Mitchell, and Stinson 2004). More formally, Valen Johnson (2003) provides an outstanding and alarming compilation of *controlled* studies demonstrating conclusively how favorable student assessment survey results are dramatically *induced* by simply inflating student grades.

Many faculty vehemently deny using self-reporting assessment surveys: "We instead use evaluations that ask whether the course objectives from the syllabi were met!" But this is no more than asking the students, "Do you think you learned a lot about this and that?" One is still getting a self-report for an answer. At the institutional level, my university continues to have its juniors take "student perception" surveys annually, to indicate the degree to which their general education courses have contributed to the development of their cognitive abilities to, *e.g.*, think objectively; develop an open mind; think and reason creatively; problem solve; see the interrelationships among ideas, concepts, and theories; and draw conclusions after weighing evidence, facts, and ideas. And a look at Pascarella and Terenzini's (2005) recently updated précis of post-secondary educational research

confirms that a sizable portion of the literature still involves the use of student self-reports to estimate CT skills development.

The CT Disposition Survey

Since a *dispositional* aspect is the focus of many reigning definitions of CT, it's only natural that an OA instrument is offered to measure for it. The *California Critical Thinking Disposition Inventory* (Facione and Facione 1992) is designed to measure the student's inclination to exercise the following set of intellectual virtues: truth seeking, open-mindedness, analyticity, systematicity, critical thinking self-confidence, inquisitiveness, and maturity of judgment.

The *CCTDI*'s seventy-five questions are designed to detect apathy and relativism with respect to truth and rationality, *i.e.*, whether one has or lacks the goals of seeking truth and avoiding error by means of having and demanding good reasons for one's beliefs. Here is a sample of its questions—students are to indicate, on a five-point scale, the degree to which they agree or disagree:

> I pretend to be logical, but I'm not.
>
> Even if the evidence is against me, I'll hold firm to my beliefs.
>
> When faced with a big question, I first seek all the information I can.
>
> People need reasons if they are going to disagree with another's opinion.

Here again, however, we are *not* necessarily measuring the students' *actual* dispositions to apply CT skills, but rather the students' self-reported *beliefs* about their dispositions—beliefs that can be wildly wrong, especially in these times of grade inflation and unrealistically high levels of self-esteem (Baumeister et al. 2005; Stout 2000: 43). And, of course, the students could well be lying about their CT habits, for a multitude of reasons. So the *CCTDI* has changed the *lyrics* a bit, compared to the self-reporting survey, but the tune is still the same:

> Survey: Do you think you think this way?
>
> Student: Yes, I think I think that way.
>
> Faculty: Then you tend to think that way.
>
> Administrator: Yes, they tend to think that way!

And the authors of the *CCTDI* admit, in the test's *Manual*, that

> [t]he *CCTDI* is not intended to be a measure of the person's CT ability or skill. A person may value being objective, but not be able to achieve objectivity. A person may be disposed toward approaching problems analytically and systematically, but not be adept at the CT skills required to do so. (Facione, Facione, and Giancarlo 2001: 6)

Facione, Facione, and Giancarlo admit that the correlation between one's score on the *CCTDI* and one's actually *having* CT skills, as measured by his own *California Critical Thinking Skills Test*, "has been fairly weak in most samples of college students" (Facione, Facione, and Giancarlo 2001: 7; see Facione 2000b for specific correlation figures).

The *CCTDI* is available through Insight Assessment, at a cost of approximately $7 per student, which includes the test booklets, answer forms, and scoring services. In light of its high price and its poor ability to assess the disposition to apply *actual* CT skills, then, I cannot recommend it.

Portfolios

Another popular OA tool is the portfolio—a collection of student work, written or multimedia. Let's assume, purely for the sake of argument, that a portfolio consists of quality contents, *e.g.*, analytical and argumentative essays and critical reviews. Can that collection of work be a reliable gauge of the student's acquisition of CT skills?

Well, if and only if there are no alternative explanations for the quality of that work, other than the enhanced CT skills of its author. For example, are the good materials in the portfolio the result of the law of large numbers or a very selective sample? Are they the result of so many revisions that they are functionally equivalent to having been authored by the instructor? These are just a few of the many considerations to keep in mind. So, while portfolios are not self-reporting, as are surveys, they are *self-selecting*, making them similarly suspect.

Assessment Essays

The written CT *essay test* starts us down the right track, in my estimation, and is the assessment tool of choice for those faculty who focus their curriculum more *globally* on the construction and critical review of alternative positions and arguments for and against them.

The *International Critical Thinking Test* (Paul and Elder 2001b) is one such example, in which the students critically examine a short position paper. Eighty percent of one's score is based on one's *analysis* of the target essay, while *only* 20 percent is based on one's *critical review* of it. I say "only" because, in my view, CT involves much more than reading comprehension. Its accent should at least equally be on *critique*—assessing claims and one's reasons for making them.

The student is first instructed to *analyze* the target article by detailing the following as completely as possible:

1. The *main purpose* of the article
2. The *key question* the author is addressing
3. The most *important information* in the article

4. The *main conclusion* of the article
5. The *key concepts* of the article
6. The *main assumptions* of the article
7. The *implications* of the article
8. The *point of view* of the author

Students are given a score of 0–10 points on how well they describe each of these aspects in their answers. And right there we find a challenge—to *consistently* score the essays, especially when multiple graders are used or the target articles are changed from one session to another (the choice of the target article being left entirely to the instructor).

And this leads us naturally to another problem, actually claimed by the authors to be a *feature* and not a flaw—the *ICTT* "is the perfect test to teach to" (Paul and Elder 2001a). With the target article selected by the faculty, and the student essays graded by that faculty, there is too much room here for self-fulfilling prophecy on student competence, as the faculty selects a target article virtually identical to one already studied.

Another problem with this CT test lies in the confusing redundancy of the various aspects on which the student is instructed to comment: One would hope that the *purpose* of the article is to argue for its *conclusion*. The *key question* addressed is, "What *conclusion* ought one to adopt?" The main *implication* of the article is its *conclusion*. And so on. So aspects 1, 2, 4, 5, 7, and 8 pretty much distill down to one thing, *viz.*, *identifying the conclusion*. (Paul and Elder might well have more in mind when talking about an article's "purpose" and its author's "point of view" than the author's conclusion, but such ventures into the author's ulterior motives strike me as inviting ad hominem fallacies instead of critical thinking.) And aspects 3 and 6, referring to the article's *information* and *assumptions*, merely have the student identify the *premises*, explicit and implicit respectively. Being able to track the anatomy of a position paper and the anatomy of its arguments *are* crucial CT skills, but the *ICTT*'s method for getting the students to demonstrate these abilities is too confusing.

The student next *critically reviews* the target article by answering the following questions:

1. Does the author clearly state his or her meaning?
2. Are the author's claims accurate?
3. Is the author sufficiently precise in providing details?
4. Does the author wander from his or her point?
5. Does the author address important complexities?
6. Does the author consider other relevant points of view?
7. Is the text internally consistent?
8. Is the text significant?

9. Is the author fair minded?

Notice, however, the glaring omission of the most important question, from the point of view of CT: Is the conclusion of the article adequately supported by the premises, especially relative to any alternative positions discussed in the article? Paul and Elder provide virtually no guidance for scoring this part of the test, other than that it should be done "holistically," with the following consideration in mind (ironically omitted from the students' list above): "To what extent does the student recognize the strengths and weaknesses in the reasoning found in the writing sample?"

Lastly, the $1000 fee for a site license to use the *ICTT* is prohibitive, *especially* when *the faculty* has to supply the target article. All things considered, there are so many problems with this test, I can only recommend against it.

A much better alternative is the *Ennis-Weir Critical Thinking Essay Test* (Ennis and Weir 1985). This forty-minute test has a single, standardized, eight-paragraph target article (a mock letter to the editor, arguing for the prohibition of overnight street parking in the neighborhood). Having a single target article is necessary in order to do extended studies using pre- and post-course testing and comparative studies across curricula, to demonstrate curricular effectiveness. And the target article is a *good* one, with plenty of formal and informal reasoning flaws, and a couple such virtues, for the students to reconstruct and discuss in their critical reviews.

One critic (whose name escapes me) found the test guilty of "cultural bias," claiming that the topic of on/off street parking was unintelligible to many of his/her inner-city students. This surprises me, however, since urban students would seem *especially* familiar with this issue. Wouldn't it be *rural* students who can't "relate" to the topic? And even that's a strain, since students simply aren't that naive, often facing this very issue while at college. So, I fear that this critic's type of qualm would quickly range over *any* topic whatsoever, thereby unjustifiably chastising *all* forms of substantive assessment.

The *Ennis-Weir*'s directions for the students are simple and clear:

> Read the letter to the editor of the Moorburg newspaper. Consider it paragraph by paragraph and as a total argument. Then write a letter to the editor in response to this one. For each paragraph in the letter you are about to read, write a paragraph in reply telling whether you believe the thinking good or bad. Also write a closing paragraph about the total argument. Defend your judgments with reasons.

In writing their critical reviews, students are thereby tested on their ability to identify and assess arguments for cogency and identify informal fallacies (not necessarily by name, however). The answer key supplied by Ennis and Weir in their test manual (1985) is very

thorough, clear, and accurate, making it easier to reach a respectable degree of consistency among multiple graders (albeit, never *completely* alleviating this inherent challenge with *any* essay test).

From among the modest concerns about reliability and validity with any essay test, I would just like to remind you of one here, and that is *experimenter's expectation*. Graders need to give the students' essays a *blind reading*, to avoid falsely attributing improved critical thinking skills to the post-curriculum students. But blind readings are difficult to set up for lone instructors.

Despite these cautions, the *Ennis-Weir* lends itself well to courses that focus on developing the *global* CT skills involved in rationally adopting, defending, and critically reviewing beliefs, values, and actions. The *Ennis-Weir* test manual kit is no longer formally published, but a free PDF version is available from Robert Ennis, so the price is right. Ennis has also compiled extensive data on the test. One such interesting finding is that students taking a *symbolic logic* course show *no* statistically significant improvement in their CT skills as measured by this test; which, when you think about it, is what you *would* expect—formal logic being but a *tiny* portion of CT (Ennis 1998).

It's rather ironic that Facione, Facione, and Giancarlo (2001: 8) claim that the *Ennis-Weir* is "not designed to challenge one's CT dispositions." Norris and Ennis, however, correctly point out (1989: 80–81) that the *Ennis-Weir* "tests for some critical thinking dispositions" by giving the student an opportunity to appraise an argument, formulate written criticisms in response, while avoiding various fallacies, thereby exercising one's CT dispositions (thereby representing them more, in my estimation, than Facione's *CCTDI survey* does). This was confirmed by Norris (2003), who found that the *Ennis-Weir* truly tested for students' disposition to apply CT skills: Students' scores on the test *improved* when they were supplied suggestions as to what to keep in mind as they critically reviewed the mock letter to the editor. Scores on objective, multiple-choice tests, on the other hand, did *not* so improve, indicating that while such tests may prompt and accurately measure CT skills, they may not be measuring one's *unprompted disposition* to put them to use.

Another genre of writing assessment tools—the *questionnaire*—has also been offered for measuring students' ability to *reflectively judge* positions on *open-ended issues*, or what are called *ill-structured problems*. Students write about (or indicate) their beliefs, and their reasons for having them, regarding various topics raised by prompts such as the following.

> Some researchers contend that alcoholism is due, at least in part, to genetic factors. They often refer to a number of family and twin studies to support this contention. Other researchers, however, do not think that alcoholism is in any way inherited. They claim that alcoholism is psychologically determined. They also claim that the reason that several members of the same family often suffer from alcoholism is due to the fact that they share family experiences, socio-economic status, or employment. (Kitchener, King, and Wood 2000)

This example is from the *Reasoning about Current Issues* test, which is available as an online exam ($2 per student). Other such questionnaires are *The Measure of Epistemological Reflection* and *The Measure of Intellectual Development*—for details regarding authors and citations, please see (Pascarella and Terenzini 2005).

But rather than getting down to the real business of assessing students' CT skills, these tests simply rate students, using, for example, William Perry's (1970) stages of intellectual and ethical development—beginning with dogmatism or blind obedience to authority, proceeding through an all-opinions-are-equal brand of relativism, and progressing towards a more contextualist approach in which the student thinks that some positions are indeed more reasonable to believe than others, depending on which opinions have the best evidence in their favor and the fewest criticisms plaguing them. Unfortunately, just *categorizing* a student as residing at a certain "stage" (1–7) gives one very little information about the student's acquisition of *specific* CT skills.

Recently the *Collegiate Learning Assessment* (2004) test, by the Council for Aid to Education, has become available as an online writing assessment tool. Because of its high cost, beginning at $6,300 for 100 students, this test is more appropriate for use at the institutional level than in the classroom; but it *is* an impressive test, so I would recommend that you try to take advantage of it if your university has already purchased a license to use it.

In one form of the *CLA*, the student has forty-five and then thirty minutes to respond to two prompts.

> The *make-an-argument* prompt presents an opinion on an issue and asks the students to address the issue from any perspective they wish, so long as they provide relevant reasons and examples to explain and support their views on topics such as: "Public figures such as actors, politicians, and athletes should expect people to be interested in their private lives. When they seek a public role, they should expect that they will lose at least some of their privacy."
>
> The *break-an-argument* prompt requires students to critique an argument by discussing how well reasoned they find it; they must do so by considering the soundness of the argument's logic (rather than agree or disagree with the position presented). An example prompt is:
>
> > The following is from an editorial in the Midvale Observer, a local newspaper. "Ever since the 1950's, when television sets began to appear in the average home, the rate of crimes committed by teenagers in the country of Alta has steadily increased. This increase in teenage crime parallels the increase in violence shown on television. Accord-

> ing to several national studies, even very young children who watch a great number of television shows featuring violent scenes display more violent behavior within their home environment than do children who do not watch violent shows. Furthermore, in a survey conducted by the Observer, over 90 percent of the respondents were parents who indicated that primetime television programs between 7 and 9 p.m. should show less violence. Therefore, in order to lower the rate of teenage crime in Alta, television viewers should demand that television programmers reduce the amount of violence shown during prime time." (http://www.cae.org/content/pro_collegiate_sample_measures.htm)

What is particularly fascinating about this assessment of students' abilities to construct and critically review arguments is that the students' responses are *scored by computer*, which the authors of the *CLA* claim is nearly as reliable as graders in the flesh—with a strength of correlation of .78 and .85 respectively (Benjamin and Chun 2003).

The other form of the *CLA* is a ninety-minute test in which students must write an argumentative proposal for solving an issue presented in the prompt, using various resources they have access to onscreen. Here's an example:

> You are the assistant to Pat Williams, the president of DynaTech, a company that makes precision electronic instruments and navigational equipment. Sally Evans, a member of DynaTech's sales force, recommended that DynaTech buy a small private plane (a SwiftAir 235) that she and other members of the sales force could use to visit customers. Pat was about to approve the purchase when there was an accident involving a SwiftAir 235. You are provided with the following documentation [by means of links]:
>
> 1: Newspaper articles about the accident
> 2: Federal Accident Report on in-flight breakups in single engine planes
> 3: Pat's e-mail to you & Sally's e-mail to Pat
> 4: Charts on SwiftAir's performance characteristics
> 5: Amateur Pilot article comparing SwiftAir 235 to similar planes
> 6: Pictures and description of SwiftAir Models 180 and 235
>
> Please prepare a memo that addresses several questions, including what data support or refute the claim that the type of wing on the SwiftAir 235 leads to more in-flight breakups, what other factors might have contributed to the accident and should be taken into account, and your overall recommendation about whether or not DynaTech should purchase the plane. (http://www.cae.org/content/pro_collegiate_sample_measures.htm)

This is a rather nice exercise to determine students' ability to judge relevant information, construct criticisms of alternative views, and construct cogent arguments for their own positions. Understandably, this portion of the *CLA* is *not* scored automatically, but is instead done by professional graders (although, according to Shavelson and Haung 2006, it, too, may become computer-scored soon).

Besides the prohibitive cost of the *CLA*, the fact that instructors have no access to the scoring criteria is problematic. It is difficult to determine exactly which CT skills are being looked for in the student responses. This question becomes especially pressing as the students have been told to "address the issue from any perspective—no answer is right," which invites a rather sophistical approach in which the goal is simply to have a response—*any* response—and not necessarily a *cogent* one. When I pressed Marc Chun on this, during a Web conference, he confirmed this, saying that the graders are just looking for students to give *some* reason or other for taking any side. But this ignores the distinction between *justifying* one's opinion and merely *rationalizing* it—a distinction ignored at CT's peril.

In spirit, the *CLA* most closely matches the open-ended test of students' CT skills in solving real-life's fuzzy problems which many are looking for, *e.g.*, Halpern (2003). All things considered, then, the *CLA* has understandably become a prominent *writing* OA test at the *institutional* level, for measuring the more *global* CT skills of critically reviewing and arguing for one's beliefs, values, and actions. (And the Spellings Commission on the Future of Higher Education cited the *CLA* as its lead example of a possible *nationally required* standardize CT test [Spellings Commission on the Future of Higher Education 2006].) But for the *classroom* level, I would recommend the cheaper, easier to use *Ennis-Weir*.

Objective Assessment Tests

Let's turn now to *objective* OA tests and their abilities to measure the acquisition of more fine-grained CT skills. Some critics are especially skeptical of this means of assessment, *e.g.*, Govier (1987) and McPeck (1981, 1991). In a sense, they are asking, "How could any little multiple-choice test determine whether one has the CT skills called upon to deal with life's multitude of questions, issues, and problems?!" The answer, of course, is that it can't. But, then, no one ever *claimed* it could—straw man fallacy. And the fallacy of false dichotomy to boot—just because we can't test *exhaustively* for CT skills doesn't mean we shouldn't do so in the best way we can.

Another complaint with these tests is that their questions are too artificial and subject to interpretation, which is always a function of context, background knowledge, assumptions, *etc.*, *etc.* For each test question, then, there are just too many *possibly* correct answers (Groarke 2005). But this criticism tends to slip the leash and range over *all* testing, in its worry about mere *logically* possible ways that a reader could construe a question. Ambiguous questions are a bad thing

(as we shall see), but not *all* possible ambiguities are equal—and so we can't rush to the conclusion that objective CT tests are unworkable.

There are currently four multiple-choice CT assessment tests that dominate the university setting. The oldest is the *Watson-Glaser Critical Thinking Appraisal Test* (Watson and Glaser 1980, 1994), first authored in the late 1930s. It takes approximately fifty minutes to complete, comes in two analogous versions (Forms A and B), and consists of eighty questions, evenly distributed among the following categories:

1. *Inference*: Whether the conclusion is "definitely true," "probably true," "probably false," or "definitely false," based on the premise, or whether there is "insufficient data."
2. *Assumptions*: Whether an assumption is necessary or not.
3. *Deduction*: Whether the conclusion is a logical implication of the premise.
4. *Interpretation*: Whether the conclusion "logically follows beyond a reasonable doubt."
5. *Evaluation*: Whether an argument is "strong" or "weak."

While the *Watson-Glaser* test is generally focused on appropriate topics, it lacks any attempt to assess students' ability to identify informal fallacies. (Please note that I am *not* requiring that students identify informal fallacies *by name*; only that students be able to recognize that the argument in question is fallacious; *e.g.*, students should correctly identify that someone is illegitimately "using a word in two different ways," but need not know that this is called "equivocation.") Among the test's other notable omissions are arguments by analogy, formal fallacies, definitions, and inductive reasoning (Govier 1987: 256).

A serious problem is created by the unfortunate wording that the *Watson-Glaser* test uses in its directions: The student is told to mark the answer "T if you think the inference is definitely TRUE; that it properly follows beyond a reasonable doubt from the statement of facts given." This is ambiguous. If the "inference is definitely TRUE," then one should judge it as a *deductive* argument; if the conclusion merely has to be "beyond a reasonable doubt," however, one should judge it as an *inductive* argument. This ambiguity haunts two sections of the test. (And calling an *inference* and an *argument*, as opposed to a *statement*, "true," can only *add* to the confusion.) This might explain why Govier (1987: 257–58) finds that the *Watson-Glaser* doesn't consistently honor the distinction between *logical* and *conversational* implication in determining answers regarding which conclusions and assumptions are implied in the questions.

Another disadvantage of the *Watson-Glaser* is that 80 percent of its questions have two-option answers, just like a true-false test does. This increases the probability of lucky guesses and, hence, decreases

the test's capacity to detect actual improvement in CT skills by means of pre- and post-testing.

Test materials and online testing are available through Harcourt Assessment, at a cost of approximately $12 per student. In light of its rather high price and the criticisms discussed above, I cannot recommend the *Watson-Glaser*.

A better alternative is the *Cornell Critical Thinking Test Level Z* (Ennis and Millman 1985). (The *Cornell Critical Thinking Test Level Y* is targeted for grades 5–12.) Ennis's understanding of the constitutive elements of CT is well defined (Ennis 2002), so his test for those elements is correspondingly quite good. Here is the distribution of test questions and their topics:

1–10	Deductive reasoning
11–21	Informal fallacy identification
22–25	Acceptability of premises
26–38	Inductive reasoning: conclusion identification
39–42	Inductive reasoning: premise identification
43–46	Definition and implicit premise identification
47–52	Implicit premise identification

Included with the test is a very thorough and recently updated manual (Ennis, Millman, and Tomko 2004), with an annotated answer key and data concerning the test's history of use and its reliability and validity. The *Cornell Test* is readily available and easy to administer, score, and analyze. All materials or software are available from The Critical Thinking Company, and cost approximately $2–3 per test. Based on a two-year study using the *Cornell Test*, I am reasonably satisfied with the test's focus and accuracy (Possin 2004). However, the *Cornell Test* does have the following limitations.

Inductive reasoning is the topic of 33 percent of the questions; so for those of us not proportionally covering that aspect of CT, focusing instead on the analysis and critical review of position papers, the *Cornell Test* is not as in sync with our curriculum as other tests may be. One might also regret its failure to test for wider competence in identifying informal fallacies, arguing by analogy, and judging the relevance of premises (Govier 1987: 266). One must keep in mind, however, that there is only so much Ennis can test for in fifty-two questions.

A general concern may arise with the test's use of only three-option answers. With fewer options, correct guesses may well have a higher probability than they should, and thus the test may not be as sensitive at detecting the actual enhancement of CT skills as it could be with a four- or five-option format. This criticism, however, raises an interesting controversy that I will address in a moment.

Lastly, there are a few problems with specific questions. For example, Govier (1987: 263–64) argues persuasively that questions 32 and

45 both have two reasonable answers. I found a similar problem with two other questions: In the last section of the test, the task is to select the answer "that is most probably the *unstated assumption*" regarding explanations. But the questions are ambiguous as to whether one is being asked to identify *an* explanation being assumed by the speaker in the question or *the* explanatory hypothesis being assumed by the speaker. Here is question #50, to illustrate the problem:

> MR. ALGAN: The explanation of the misbehavior of Gallton's present-day crop of youngsters is a simple one. These children have been severely punished at some time or other. That's the trouble.
>
> A. Children who have been severely punished misbehave.
>
> B. Children who misbehave have been severely punished at some time.
>
> C. Children who haven't been severely punished behave properly.

The correct answer, according to the *Test Manual*, is "A." Admittedly, "A" would make being severely punished *an* explanation of misbehavior, but not *the* explanation of the misbehavior (which is what Mr. Algan is claiming to give). In order to be *the* explanation, it would *also* have to be the case that the children probably wouldn't exhibit this misbehavior without having been severely punished, *viz.*, "C." And answer "B" is the contrapositive of "C." So, the correct answer *should* be "A," "B," *and* "C." And similarly for question #51.

In correspondence, Ennis responded to this by saying, "I agree that singular causes often are necessary for the effect, but not necessarily so. They are not necessary when the effect is linked overdetermined." Ennis is referring here to a type of causal situation made famous by Michael Scriven (1966), in which, had the original cause not occurred, another back-up cause, perhaps even triggered by the failure of the original, would have brought about the effect. But I don't think this resolves the issue. First, we have to remember that the test tells us to find the person's *most probable* assumption, and the average person doesn't likely have overdetermination in mind as an explanatory possibility. Second, Mr. Algan says the explanation is a *simple* one, and overdetermined cases are not very simple. And lastly, if overdetermination were really a believed option for Mr. Algan—say he believed kids were genetically predisposed to misbehave anyway—he would *not* likely be claiming to know that severe punishment is *the* explanation. No, the best explanation for why he cites it is because he *also* thinks that, in this case, but for the severe punishment, the misbehavior wouldn't likely have occurred.

Overall, however, the *Cornell CT Test Level Z* is well-constructed and has a well-documented history and thereby gets my recommendation, *especially* for those faculty who include the study of inductive reasoning in their CT curriculum.

The *Cornell Test*'s main competitor is the *California Critical Thinking Skills Test* (Facione 1990, 1992, 2000a), which has become a very popular CT assessment test internationally. It consists of thirty-four multiple-choice (both four- and five-option) questions, taking forty-five minutes to complete. There are three versions of the *CCTST*, Forms A, B, and 2000. All three versions focus on the following overlapping categories of CT skills (Facione et al. 2002):

- *Analysis*: Identifying arguments, their conclusions and premises.
- *Evaluation*: Assessing the cogency and relative strength of inferences.
- *Inferences*: Drawing conclusions and identifying premises.
- *Deductive Reasoning*: Judging the validity of arguments.
- *Inductive Reasoning*: Judging if the conclusion is probable, given the premises.

Conspicuously absent from the *CCTST* is the assessment of one's ability to judge the acceptability of premises and identify informal fallacies. Conspicuously *added* to the newest Form 2000 are questions concerning two flowcharts and a pie chart. At first I thought that this update made the *CCTST* less of a CT test and too much of a *reading comprehension* test; but Peter Facione, in correspondence, changed my mind on this by pointing out that quite often now our inferences *are* based on graphical representations, instead of premises in prose, so students *should* be tested on this competency too.

That the *CCTST*'s number of answer options is more than three is construed by many as a virtue—the more options, the less lucky guessing is rewarded. But in an impressive review of the literature, Michael Rodriguez (2005) disputes this, finding that three-option answers are often as effective as four- or five-option answers, because the extra items on the latter are too often merely "throw away" distracters—so obviously incorrect that examinees ignore them anyway. While this may be true generally, I don't find it to be the case with the *CCTST*—its distracters are quite tempting. However, another point by Rodriguez is certainly true: Questions with more answer options take longer to answer, which would explain why the *CCTST* has thirty-four test questions (with an allotted time of forty-five minutes) and the *Cornell Test* has fifty-two items (with an allotted time of fifty minutes). This enables the fewer-option *Cornell* to increase its content coverage, without sacrificing its validity.

Sometimes the wording of questions in the *CCTST* gets a bit too tricky, and the student is tested less on thinking critically and more on reading meticulously. This is especially true for Forms A and 2000

(which is why I chose Form B for use in my CT courses). As an example, on Form 2000, one question reads as follows:

> Passage: "The microorganisms in this pond are of the kind which generally reproduce only in water with a temperature above the freezing point. Now it's winter time and this pond is solid ice. So, if there are any microorganisms of the kind we are researching in the pond, they aren't reproducing right now." Assuming all the supporting statements are true, the conclusion of this passage. . . .

The accepted answer is that the conclusion "is probably accurate, but may be inaccurate." Indeed, that is *technically* right—the operative word in the passage being "generally," which turned this into an *inductive* argument. Admittedly, "generally" is an important qualifier that a person should notice, but the test should not make it unusually difficult to do so, which it did by having this question immediately appear after numerous *deductive* exercises; so one is rather primed to focus on the word "only" and infer, by modus tollens, that the conclusion in the passage "could not be inaccurate." Moreover, the accepted answer expects one to entertain the possibility that the microorganisms are reproducing *in solid ice*—a physical impossibility.

For each Form of the *CCTST*, there can be reasonable disputes about the occasional particular item (albeit not enough to justify dismissing any of the tests as unworkable). For example, this item from Form 2000 seems to have multiple reasonable answers.

> A friend who does not work with you tells you, "Setting aside the union contract for a moment, there is sufficient reason for firing your assistant. He has lied. He is disorganized and loses important things. He did not even check with you about sending the package late, once he found it." The friend's reasoning is
>
> A= poor, because the friend does not know the circumstances of work in your office.
>
> B= poor, because the friend has not given the assistant the chance to defend himself.
>
> C= good, because the assistant's poor work has hurt your business and your reputation.
>
> D= good, because the assistant has performed in exactly these substandard ways.

The accepted answer is D. But this strikes me as begging the question that the friend's descriptions are accurate. C gets ruled out because there was no evidence given that you were hurt yet by the assistant's behavior. Answers A and B both seem reasonable though: How informed can this friend be, when not even at the workplace, especially regarding possibly extenuating circumstances surrounding the assistant's *assumed* misdeeds?

Whichever Form you decide to adopt, the *reigning* advice has been to stick with it, because, as much as, *e.g.*, Forms A and B were *designed* to be analogous, groups of students in a controlled study were found

to produce statistically significantly different *mean* scores, along with numerous statistically significantly different mean *item* scores, all supposedly indicating that Form B is *slightly* more difficult (Jacobs 1999). However, in an analogous study, I found similarly different *mean* scores and mean *item* scores in groups taking *identical* tests (Possin 2005), indicating that Jacobs's evidence for the nonequivalence of Forms A and B is much weaker than has been thought.

The *CCTST* does have a rather serious drawback for classroom use, in my estimation. Its publisher, Insight Assessment, no longer provides answer keys for any of its tests; so one *must* use its answer sheets and scoring services, all at considerable cost and inconvenience—approximately $7 per test for either paper or online testing. To use the online testing service, one must download and install a small computer applet, *The CalPress Online Testing Tool*, which allows students Internet access to the tests and their scoring. I found the PC version of the *Testing Tool* worked well, but the Mac version did not. Overall, however, the *CCTST* is a very respectable assessment test, in terms of content; and it is widely adopted, so there are already much data and research available involving comparative studies among Forms A, B, and 2000 and other objective assessment tests (see Facione et al. 2002).

The most expensive *objective* CT assessment test, of the four I review here, is the *Critical Thinking* module of the *Collegiate Assessment of Academic Proficiency*. The *CAAP* is used by many universities to assess the effectiveness of their general education programs. Juniors are given one or two of its modules, in writing, reading, science reasoning, math, or CT. The *CAAP* is offered by the creators of the *ACT*. One must purchase answer sheets and scoring services with the test booklets, at a high cost of around $13 per student. The regulatory procedures for administering the *CAAP-CT* are so strict and secretive that they are not worth the bother, *unless* your university is already using the *CAAP* and you can dovetail your classes into it at no charge to your department. Under *those* conditions, I would deem it worth a try, because the *CAAP-CT*'s *content* is adequate.

The test consists of four short (six- to twelve-paragraph) passages—one position paper and three debates—with eight four-option questions asked about each passage. The CT skills tested for involve identifying the following: assumptions, conclusions, premises, formal flaws in arguments, logical implications, logical inconsistencies, and counterexamples. A crucial omission here, in my opinion, is the identification of *informal* flaws in arguments.

Another problem with the *CAAP-CT* is that two of the eight questions for each passage merely test reading comprehension; for example, "What was A's complaint about the current administration's policies?" For this, the student simply rereads the first paragraph of the passage to

find the answer stated. So that means 25 percent of the test is dedicated less to CT skills and more to mere reading comprehension, and that's just too much (albeit, *e.g.*, Halpern [1993] includes reading comprehension, as well as using mnemonic devices and spatial representations, in her list of CT skills).

And lastly, I found an objectionable political statement intimated in one of the passages, in which the character arguing *against* tax exemptions for religious institutions is named "Sinning," while the character arguing *for* tax exemptions is named "Saintly." This is quite unprofessional—an actual case of "cultural bias"—on the part of ACT Inc., and I can only hope that this passage has since been retired from the current two tests in rotation.

Conclusions and Implications

I hope that I've provided you with some helpful resources and alternatives for when you get the call for your "assessment plans, processes, and measures" with respect to how your courses or programs are enhancing students' CT skills.

Some *objective* CT assessment tests appear to be fairly accurate and affordable for measuring students' acquisition of specific *core* CT skills. Interestingly, these tests indicate that *dedicated CT courses are most effective at enhancing those skills*, with *computer-assisted* courses showing especially impressive results (Hitchcock 2004, Possin 2004, van Gelder 2001). Other philosophy courses, *e.g.*, intro and even *logic* courses, have shown no statistically significantly better results on these tests than other university courses generally (Hitchcock 2004), or only slightly better results (.25 SD v .12 SD) (van Gelder 2007). Neither do students' majors (Pascarella and Terenzini 2005) show statistically significant differences in particular. So much, then, for leaving the task of enhancing CT skills in our students to "immersion" and "critical thinking across the curriculum." Let me note an exception, however: by adding a *separate*, *dedicated*, *generic* CT curriculum to his general psychology course, Tom Solon (2006) demonstrated impressive improvement in his students' CT skills, using the *Cornell Test*. But this is *not* what people *usually* do when they claim to incorporate CT into their classroom (or ballroom?).

And some CT assessment *writing* tests appear to measure fairly well the enhancement of the more *global* CT skills involved in rationally adopting and critically reviewing positions and their arguments and criticisms. We should remember that symbolic logic courses strike out here too, their students showing no improvement in CT skills by means of the *Ennis-Weir* test (Ennis 1998)—which is yet *another* reason why

philosophy departments need to offer *dedicated CT courses* and not just think that their logic courses are doing the job.

To get the most complete measure of CT skills improvement, one would *ideally* administer *both* types of assessment tests, pre- and post-course. This is exactly what Donald Hatcher (2006) has done. By melding its *dedicated, generic* CT curriculum and its Freshman Comp curriculum, Baker University created a writing intensive three-course program, and was able to demonstrate respectable gains with the CCTST, .57 of a standard deviation (SD), and impressive gains with the *Ennis-Weir*, 1.47 SD—statistically significant improvement in the post-course test groups, albeit not to the same extent with both types of tests. This has been my experience too, based on my use of the *Cornell Test* and numerous "Anatomy of a Position Paper" assignments: I found improvement on both post-CT tests, though not necessarily with respect to the *same students*. This further illustrates how CT involves *both* core *specific* skills of argument identification and assessment *and* more *global* skills involved in the critical analysis of issues. It reminds us again that *different people* acquire those *different CT skills* at *different rates*, although *both categories of skills are equally important.*

Perhaps CT assessment *essay* tests *would* reveal that general philosophy courses enhance the more *global* critical thinking skills involved in rationally adopting and critically reviewing positions and their arguments, just as dedicated CT courses demonstrably enhance those skills. One would *hope* that such CT skills are cultivated in most philosophy courses, especially in ones requiring writing assignments. We like to *claim* that we, as a *discipline*, are enhancing students' CT skills, *but do we have the evidence to back up that claim?*

Postscript—Doing OA

You've selected your OA test; now, how should you administer it and what should you do with the data? Here are a few suggestions that might help, the first of which is to befriend a statistician, for better assistance than I could possibly provide here.

There are two very popular methods for evaluating your test results. One is the *paired t-test for individuals* and the other is the *t-test for independent samples*. With the former method, you give the same test to the same students before and after the course and calculate the mean difference between their *individual* pre- and post-course scores. With the latter method, the difference between the mean score of the pre-course *group* and the mean score of the post-course *group* is calculated.

The paired t-test for individuals is considered the more accurate gauge, because it focuses more directly on the individuals' improvement. For example, with a t-test for independent samples, students

who ultimately drop the course cull themselves from the post-course group but not from the pre-course group, where they may have lowered its mean score. And even if measures are taken so that the members of the pre- and post-course test groups are identical, the paired t-test for *individuals* will more exactly calculate the difference between the *groups*, because of its inherently smaller margins of error.

If you administer only the post-course test, because you do not want the students taking the same test twice, so as to avoid *test carryover*, you have the added challenge of finding a *control* group similar in relevant respects to the group of students you're testing. For example, in a *Cornell Test* study involving my CT course (Possin 2004), I used students attending their first day of an introductory philosophy course as my control group. Test carryover is a concern with *essay* tests such as the *Ennis-Weir*, but, as I only later learned, not with *objective* tests (Facione et al. 2002). (Perhaps this partially explains why the majority of OA studies I'm aware of have used t-tests for independent *groups* instead of the more accurate paired t-tests for individuals.)

So, what difference in test scores would warrant your claim to have *made* a difference in the CT skills of your students? To help me answer this question, here are a few more details about my 2004 study using a t-test for independent groups taking the *Cornell Test*: The mean score for the (control) group (N=129) *not* taking my CT course was 26.98, with a standard deviation (SD) of 4.53. The mean score of the post-CT group (N=416) was 30.42, with an SD of 5.15. So the students completing my CT curriculum (Possin 2002a) scored an average of 3.44 points higher, this difference being statistically significant ($p < 0.0001$). Another way to state this is to say that the *size of effect* for the course was .76 of an SD. This *effect size* is technically called "Cohen's *d*," but all we need to know is that it is a rather nice unit for measuring differences between pre- and post-course results, even when using various sample sizes, various tests, and various methods for evaluating the score data. To calculate effect size, take the mean improvement in scores in the post-course group and divide it by either the *weighted* ("pooled") standard deviation for *both* groups or the standard deviation for the pre-course or control group (the latter calculation being easier and often used).

Once you've been testing for a while, you might begin noticing some surprising results concerning the cultivation of CT skills in your students. For instance, I found that increasing class size does *not* seem to matter adversely, when one's CT curriculum is computer-assisted, confirming Hitchcock (2004). When my curriculum was taught by a sabbatical replacement, I also confirmed Don Hatcher's finding (Hitchcock 2004) that effect size tends to be lower with less teaching experience. And shortening the academic calendar, thereby sacrificing course content, similarly lowers effect size—yes, Academic Calendar Committee, it *does* matter when one semester is markedly shorter than another.

So what *is* a *respectable* effect size to achieve on your CT assessment tests? Well, according to research by Pascarella and Terenzini (1991), the effect size of four years of undergraduate education on students' CT skills was one SD—measured by giving entering freshmen and exiting seniors CT assessment tests such as those I've discussed. More recently, however, these same authors (2005) found that the effect size of the baccalaureate on students' CT skills has shrunk somewhat, but still remains at least above .55 SD, with the greatest effect size occurring during freshman year (.44 SD). Most recently, Tim van Gelder (2007) found it remaining at approximately one SD, based on his meta-analysis of 52 studies involving only *objective* CT assessment tests. The effect size of a *dedicated* CT course is .23 SD, according to Pascarella and Terenzini (2005), but they are cautious about this claim due the lack of consensus about what constitutes CT. According to van Gelder (2007), its effect size is .34 SD when taught by philosophy faculty and .4 SD when taught by others. So, with an effect size of .76 SD, I guess I did pretty well, enhancing my students' CT skills in one semester what otherwise would have taken them approximately three or four years of university coursework to accomplish.

Reasons why Pascarella and Terenzini found a recent decline in the baccalaureate's effect size on CT skills might stem from the broader range of OA tests in their meta-analysis. All they are willing to say is that "on an absolute standard, not all college graduates are proficient critical thinkers" (Pascarella and Terenzini 2005: 205). Indeed. For instance, among the many bleak findings in a study by the American Institutes for Research (2006) is that more than 50 percent of graduating seniors at four-year colleges (and more than 75 percent of the graduating students at two-year colleges) lack the skills to even compare credit card offers with different interest rates or follow the arguments in a newspaper editorial. And let's not forget the AACU's finding (2005) that only 6 percent of graduating college seniors are proficient at CT. Looks like we've got some work to do! Best of luck, then, at improving the CT skills of *your* students and at using some of these OA tests to document it.

Acknowledgments

Many thanks go to David Hitchcock, for providing test samples, advice, and critical comments on earlier thoughts on this topic. I also want to express my appreciation to Robert Ennis, Don Hatcher, and Joyce Quella, for all their help; and especially to an anonymous reviewer for this journal, who provided many excellent criticisms and suggestions.

Bibliography

ACT Inc. (2000). *The Collegiate Assessment of Academic Proficiency—Critical Thinking Test.* Iowa City, IA: ACT Inc.

American Institutes for Research. 2006. *The National Survey of America's College Students.* Available at http://www.air.org.

American Philosophical Association. 1990. *Critical Thinking: A Statement of Expert Consensus for Purposes of Educational Assessment and Instruction* (also known as *The Delphi Report*) (Millbrae, Calif: California Academic Press). Peter Facione's Executive Summary, is available as a PDF at http://www.insightassessment.com/dex.html.

_____ . 2001. *APA Statement on Outcomes Assessment.* Available at http://www.apa.udel.edu/apa/governance/statements/outcomes.html.

Association for Informal Logic and Critical Thinking. 2007. *Resolution on the Spellings Commission Report.* Available at http://ailact.mcmaster.ca/index.html.

Association of American Colleges and Universities. 2005. *Liberal Education Outcomes: A Preliminary Report on Student Achievement in College.* Available at http://www.aacu.org/press_room/press_releases/2005/Outcomes.cfm.

Baumeister, R., J. Campbell, J. Krueger, and K. Vohs. 2005. "Exploding the Self-Esteem Myth," *Scientific American* 292:1: 84–91.

Benjamin, R., and M. Chun. 2003. "A New Field of Dreams: The Collegiate Learning Assessment Project," *Peer Review* (Summer): 26–29.

Bloom, B. 1956. *Taxonomy of Education Objectives* (New York: Longmans Green).

Council for Aid to Education. 2004. *The Collegiate Learning Assessment Test.* Available at http://www.cae.org/content/pro_collegiate.htm.

Dresner, J. 2004. "Grade Inflation: Why It's a Nightmare," *History News Network.* Available at http://hnn.us/articles/6591.html.

Ennis, R. 1984. "Problems in Testing IL/CT/Reasoning Ability," *Informal Logic* 6:1: 3–9.

_____ . 1998. *Manual Supplement for* The Ennis-Weir Critical Thinking Essay Test. Available upon request from the author.

_____ . 2002. "An Outline of Goals for a Critical Thinking Curriculum and Its Assessment." Available at http://faculty.ed.uiuc.edu/rhennis/outlinegoalsctcurassess3.html.

_____ . 2008. "Nationwide Testing of Critical Thinking for Higher Education:Vigilance Required," *Teaching Philosophy* 31:1 (March): 1–26.

Ennis, R., and J. Millman. 1985. *The Cornell Critical Thinking Test Level Z* (Pacific Grove, Calif.: Midwest Publications).

Ennis, R., J. Millman, and T. N. Tomko. 2004. *Manual, The Cornell Critical Thinking Test Level Z,* 4th ed. (Seaside, Calif.: The Critical Thinking Company).

Ennis, R., and E. Weir. 1985. *The Ennis-Weir Critical Thinking Essay Test: Test, Manual.* Available at http://faculty.ed.uiuc.edu/rhennis/Assessment.html.

Facione, P. 1990. *The California Critical Thinking Skills Test, Form A* (Millbrae, Calif.: California Academic Press).

_____ . 1992. *The California Critical Thinking Skills Test, Form B* (Millbrae, Calif.: California Academic Press).

_____ . 2000a. *The California Critical Thinking Skills Test, Form 2000* (Millbrae, Calif.: California Academic Press).

_____ . 2000b. "The Disposition Toward Critical Thinking: Its Character, Measurement, and Relationship to Critical Thinking Skill," *Informal Logic* 20:1: 61–84. Available at http://www.insightassessment.com.

Facione, P., and N. Facione. 1992. *The California Critical Thinking Disposition Inventory* (Millbrae, Calif.: California Academic Press).

Facione, P., N. Facione, S. Blohm, and C. Giancarlo. 2002. *Manual, The California Critical Thinking Skills Test, Form A, Form B, Form 2000* (Millbrae, Calif.: California Academic Press).

Facione, P., N. Facione, and C. Giancarlo. 2001. *The California Critical Thinking Disposition Inventory: Inventory Manual* (Millbrae, Calif.: The California Academic Press).

Felton, J., J. Mitchell, and M. Stinson. 2004. "Web-Based Student Evaluations of Professors: The Relations Between Perceived Quality, Easiness and Sexiness," *Assessment & Evaluation in Higher Education* 29:1: 91–108.

Govier, T. 1987. *Problems in Argument Analysis and Evaluation* (Dordrecht: Foris).

Gratton, C. 2001. "Common Pedagogical Weakness in Critical Thinking Textbooks and Courses," in *Proceedings of the 2001 OSSA Conference: Argumentation and Its Applications,* ed. H. Hansen, C. Tindale, A. Blair, and R. Johnson (Windsor, Ont.: Informal Logic).

Groarke, L. 2005. "What is Wrong with the *CCTST*? Critical Thinking Testing and Educational Accountability." Association for Informal Logic and Critical Thinking, Pacific meeting.

Halpern, D. F. 1993. "Assessing the Effectiveness of Critical Thinking Instruction," *The Journal of General Education* 42:4: 238–54.

_____ . 2003. "The 'How' and 'Why' of Critical Thinking Assessment," in *Critical Thinking and Reasoning: Current Research, Theory, and Practice,* ed. D. Fasko (Cresskill, N.J.: Hampton Press).

Hatcher, D. 2006. "Stand-Alone Versus Integrated Critical Thinking Courses," *The Journal of General Education* 55:3-4: 247–72.

Hatcher, D., and A. Spencer. 1993. *Reasoning and Writing: An Introduction to Critical Thinking* (Lanham, Md.: Rowman and Littlefield).

Hitchcock, D. 2004. "The Effectiveness of Computer-Assisted Instruction in Critical Thinking," *Informal Logic* 24:3 (Fall): 183–217.

Isenberg, A. 1949. "Critical Communication," *The Philosophical Review* (July): 330–44.

Jacobs, S. 1999. "The Equivalence of Forms A and B of the California Critical Thinking Skills Test," *Measurement & Evaluation in Counseling & Development* 31:4: 211–23.

Johnson, R., and A. Blair. 1987. "The Current State of Informal Logic," *Informal Logic* 9: 147–51.

_____ . 1991. "Misconceptions of Informal Logic: Reply to McPeck," *Teaching Philosophy* 14:1: 35–51.

_____ . 1994. *Logical Self-Defense* (New York: McGraw-Hill).

Johnson, V. 2003. *Grade Inflation: A Crisis in College Education* (New York: Springer-Verlag).

Kitchener, K., P. King, and P. Wood. 2000. *Reasoning about Current Issues.* Available at http://www.umich.edu/~refjudg/index.html.

McPeck, J. 1981. *Critical Thinking and Education* (New York: St. Martin's).

_____ . 1991. *Teaching Critical Thinking* (New York: Routledge).

Norris, S. 2003. "The Meaning of Critical Thinking Test Performance: The Effects of Abilities and Dispositions on Scores," in *Critical Thinking and Reasoning: Current Research, Theory, and Practice,* ed. D. Fasko (Cresskill, N.J.: Hampton Press).

Norris, S., and R. Ennis. 1989. *Evaluating Critical Thinking* (Pacific Grove, Calif.: Midwest Publications Critical Thinking Press).

Pascarella, E., and P. Terenzini. 1991. *How College Affects Students: Findings and Insights from Twenty Years of Research* (San Francisco: Jossey-Bass).

_____ . 2005. *How College Affects Students, Volume 2: A Third Decade of Research* (San Francisco: Jossey-Bass).

Paul, R., and L. Elder. 2001a. *Critical Thinking: Tools for Taking Charge of Your Learning and Your Life* (Upper Saddle River, N.J.: Prentice Hall).
_____. 2001b. *The International Critical Thinking Test* (Dillon Beach, Calif.: The International Center for the Assessment of Higher Order Thinking).
Perry, W. 1970. *Forms of Intellectual and Ethical Development in the College Years* (Troy, Mo.: Holt, Rinehart, & Winston).
Possin, K. 2002a. *Critical Thinking* (Winona, Minn.: The Critical Thinking Lab).
_____. 2002b. *Self-Defense: A Student Guide to Writing Position Papers* (Winona, Minn.: The Critical Thinking Lab).
_____. 2004. *Critical Thinking Proven Effective*. Available from the author.
_____. 2005. *The Equivalence of Forms A and B of the* California Critical Thinking Skills Test *Revisited*. Available from the author.
Rodriguez, M. C. 2005. "Three Options Are Optimal for Multiple-Choice Items: A Meta-Analysis of 80 Years of Research," *Educational Measurement: Issues & Practice* 24:2 (June): 3–13.
Scriven, M. 1966. "Defects of the Necessary Condition Analysis of Causation," in *Causation*, ed. E. Sosa and M. Tooley (Oxford: Oxford University Press).
Shavelson, R., and L. Haung. 2006. *A Brief History of Assessing Undergraduates' Learning*. Available at http://www.cae.org/content/pdf/abriefhistoryofassessingundergrad.pdf.
Sobocan, J. 2003. "Teaching Informal Logic and Critical Thinking," *Informal Logic @ 25 Symposium*. Available at http://web2.uwindsor.ca/faculty/arts/philosophy/ILat25/papers.htm.
Solon, T. 2006. "Generic Critical Thinking Infusion and Course Content Learning in Introductory Psychology," Association for Informal Logic and Critical Thinking, Central meeting.
Spellings Commission on the Future of Higher Education. 2006. *A Test of Leadership: Charting the Future of the US Higher Education*. Available at http://www.ed.gov/about/bdscomm/list/hiedfuture/reports.html.
Stout, M. 2000. *The Feel-Good Curriculum: The Dumbing Down of America's Kids in the Name of Self-Esteem* (Cambridge, Mass.: Perseus Books).
van Eemeron, F. H., R. Grootendorst, and F. Snoek Henkemes. 1996. *Fundamentals of Argumentation Theory* (Mahwah, N.J.: Lawrence Erlbaum Associates).
van Gelder, T. 2001. "How to Improve Critical Thinking Using Education Technology," *Proceedings of the 18th Annual Conference of the Australasian Society for Computers in Learning in Tertiary Education*.
_____. 2007. Re: [AILACT-D] Required College Level Thinking Test(s?). E-mail correspondence to all listserve members of the Association for Informal Logic and Critical Thinking.
Walton, D. 1989. *Informal Logic: A Handbook for Critical Argumentation* (Cambridge: Cambridge University Press).
Watson, G., and E. Glaser. 1980. *The Watson-Glaser Critical Thinking Appraisal Manual Forms A and B* (San Antonio, Tex.: The Psychological Corporation, Harcourt Brace & Company).
_____. 1994. *The Watson-Glaser Critical Thinking Appraisal Form B* (San Antonio, Tex.: The Psychological Corporation, Harcourt Brace & Company).
Weinstein, M. 1995. "Critical Thinking: Expanding the Paradigm," *Inquiry* (Fall). Available at http://www.chss.montclair.edu/inquiry/fall95/weinste.html.

Kevin Possin, Philosophy Department, Winona State University, Winona, MN 55987; kpossin@winona.edu

Teaching Modernity in Appalachia

ALEXANDRA BRADNER
Denison University

Abstract: Despite our interests in conceptual schemes, paradigms, styles of reasoning, levels of explanation, and populationist modes of theorizing, many philosophers ignore the fact that instruction occurs in situ. This paper highlights the importance of cultural location by reflecting upon the author's experience as an instructor of modernity at Marshall University, a regional state institution in Huntington, West Virginia. For many Appalachian students, issues barely tolerated by others (as part of their required history sequence) are uniquely resonant. At the same time, existing power structures—and the very real limits established by those structures—discourage Appalachian students from embracing or even entertaining the canonical themes of modernity. Immersing oneself in the regional culture, instead of bemoaning it, enables a philosophy instructor to examine modernity from both the pre- and post-modern perspectives, while also conveying to students that their education matters a great deal to the fate of the region.

WEST VIRGINIA AND HUNTINGTON:
A STATISTICAL SNAPSHOT (in approximate numbers)

- State population: 1.8 million, which is smaller than the city of Houston, but a bit larger than the city of Philadelphia.
- Ethnic composition: 95.2% checked "white" on the last census, compared with 80.4% nationally; 1.1% are foreign-born, compared to 11.1% nationally. Huntington, the state's second-largest city, has 49,533 people and is slightly more diverse. The state capital, Charleston, has approximately 51,600 people.
- West Virginians living below the poverty line of $17,463 for a family of four (according to the 2000 census): 16.3%. In Huntington alone, this figure is 24.7%. WV fares worst (51st) in the nation on this statistic. West Virginians living below the poverty line last year: 17.9%; this is the 5th highest percentage of individuals living below the poverty line in the nation.
- Median household income: $32,967 (in Huntington alone, $23,234), compared to $43,318 nationally. WV has the lowest median household income (51st) in the nation. (Of course, when thinking about all of these numbers, we might keep their international context in mind. According to the World Bank, in 2004 the gross national income per capita of the U.S. was $41,400, while Mexico's was $6,770, and Sierra Leone's $200.)
- West Virginians who have earned a bachelor's degree or higher: 14.8% (in Huntington alone, 22.4%), compared with 24.4% nationally. WV fares worst (51st) in the nation on higher education, but slightly better (48th) on the number of state residents who have completed high school.

 pp. 229–247