11/1/05

SUPPLEMENT TO THE TEST/MANUAL ENTITLED *THE ENNIS-WEIR CRITICAL THINKING ESSAY TEST*

Robert H. Ennis, Professor Emeritus, University of Illinois, UC

Rhennis@uiuc.edu

Published by

Illinois Critical Thinking Project

Department of Educational Policy Studies

University of Illinois, UC

Urbana, IL 61820

November, 2005

To be available on the Academic Web site, http://faculty.ed.uiuc.edu/rhennis

CONTENTS

TABLES

INTRODUCTION

The original test/manual (Ennis & Weir, 1985) provided the test, as well as information about its design, content-related evidence of validity, "reliability", and administration and scoring. A first supplement was prepared in 1998. The current supplement is intended to replace the first supplement and contains the data provided in that first supplement and the original test/manual, together with the considerable data acquired since then. More data are (and will be) out there, and are hereby solicited for use in preparing a subsequent supplement. But for the test itself, and information about its design, content-related evidence of validity, administration, and scoring, you are referred to the original test/manual, which is available for downloading at no cost from my academic web site, http://faculty.ed.uiuc.edu/rhennis.

The current supplement provides data (and discussion thereof) resulting from the test's use in twenty-four studies. You will find information regarding the groups involved in the studies, user norms, "reliability" (consistency) indices, experimental studies about the

effects of teaching critical thinking and other things, and relationships with other variables.

Information about these studies came to my attention either through a review of the *Social Science Citation Index* and a review of *Dissertation Abstracts International* (both from 1980 through summer of 1999), or personal communication. For various reasons, some studies were not usable.

THE NATURE OF THE GROUPS INVOLVED

In the following list, groups that are the sources of these data are designated by the letters, "EW" and a number, roughly ordered in accord with the date of the securing of the data. Each group is briefly described because the contexts of the data are relevant to their interpretation:

EW 1. Seventy-one undergraduates at a SUNY unit in upstate New York, in two groups: students at the end of a traditional introductory logic course and students at the end of an experimental two-semester sequence in writing and critical thinking called "Effective Thinking and Communicating" (Rapaport, 1979).

EW 2. Twenty-seven undergraduates in a critical thinking/informal logic course in a large midwestern state university, tested midway through the course (Ennis & Weir,1985, p. 4).

EW 3. Twenty-eight gifted eighth-grade students of English, who had received some critical thinking instruction in a suburban Chicago-area school system (Ennis & Weir, 1985, p. 4).

EW 4. Seventy-two fifth- and sixth-grade gifted students (experimental and control groups) from a large school district in Central Mississippi; the experimental group was explicitly taught standard critical thinking objectives in a seven-week portion of the curriculum (Chennault, 1989).

EW 5. Ninety-three above-average eighth-grade students in a course in library research and writing skills. (Goldberg, 1991).

EW 6. Sixty-five tenth graders (4^{th} form) in a traditional Commonwealth girls high school in Jamaica (experimental and control groups). The experimental group was explicitly taught critical thinking, using among other things a letter to an editor comparable to the Ennis-Weir Moorburg letter, and focusing on the aspects of critical thinking assessed in the Ennis-Weir test as well as some other standard "fallacies" (Sirbratthie, 1991).

EW 7. Forty-seven elementary teachers broken into three groups: those who received instruction in Richard Paul’s Remodeling of Lesson Plans with the inclusion of a thinking

frame, those who received the remodeling instruction but without the thinking frame, and those who received neither (Moreyra, 1991).

EW 8. One hundred forty-eight ninth- and eleventh-grade students taught persuasive writing one period every Monday for six weeks; the experimental group in accord with Toulmin's model of arguments, and the control group in accord with a section of the text, *Writing Persuasive Composition* (Wallace, 1992).

EW 9. Twenty-seven academically gifted students selected from twenty-one middle schools in Southern California after 40.5 hours of summer instruction in a course on "thinking-writing" (Waldron, 1992).

EW 10. Sixty gifted high school seniors in a Florida West Coast high school. The experimental group took a course in introduction to philosophy; the control group took an AP English course (Yarbrough,1992).

EW 11. Sixty-four tenth-, eleventh-, and twelfth-grade students in suburban schools near St. Johns, Newfoundland, were given three versions of the Ennis-Weir test to see whether extra hints and guidelines helped the students on the test. This was done in order to see whether critical thinking dispositions played a role in taking the Ennis-Weir test. (Norris & Hollett, 1992; Norris, 2003).

EW 12. Forty-nine heterogeneous fifth graders in rural central Pennsylvania, instructed for 30 days at 1.5 hours per day in critical thinking in English and social studies; for about half the students the critical thinking instruction was infused ("deliberate teaching of the thinking skills"), while for the others, the instruction was immersion ("thinking skills develop naturally in the context"), (Colledge, 1993).

EW 13. Thirty-three sophomores at a midsize Midwestern university enrolled in a required core course for a major in communications, about half of whom were given instruction that integrated critical thinking and deconstructive conceptual frameworks, the other half serving as a control group (Koehler, 1993).

EW 14. One hundred ninety-eight undergraduates in an educational psychology course at a large Midwestern university, who were given a number of tests to see the relation between critical thinking and certain academic and personality variables, and to check the viability of the conceptualization of critical thinking as two factors, abilities and dispositions (Taube, 1993).

EW 15. Sixty Maylay undergraduates for whom English is a second language at a large Midwestern university, pre- and post-tested in a critical thinking course. Correlations with SAT, TOEFL, ACT, TWE, and the Maylaysian Certificate of Education test are provided (Moore, 1995).

EW 16. Thirty-eight Southern-California seventh-grade students (65% minority) in a model technology program that was intended to develop critical thinking abilities. Pre-post data only (Jacks, 1994).

EW17. Twenty-nine community-college non-major chemistry students, given three laboratory exercises asking them to identify unknown chemicals. Results were checked to see whether success correlated with the Ennis-Weir test. (Morgan, 1995).

EW 18. One hundred seventy-two tenth-, eleventh-, and twelfth-grade students in St. John's, Newfoundland, who took versions of various critical thinking tests. Results were factor-analyzed (Norris, 1995).

EW19. Eighty-four students enrolled in an English-composition argumentation course in a Midwestern university (Weiner, 1996).

EW 20. Thirty-six Asian junior college women who took a one-year intensive EFL/ESL course. About half also experienced eighteen hours of critical thinking instruction. All were given the Ennis-Weir test as a post-test only (Davidson & Dunham, 1997).

EW 21. Over a period of five years, 977 freshman in a midsize Midwestern university took a one-year combined critical-thinking and writing course, yielding pre-post scores. A longitudinal comparison (freshman to senior) is provided with pre-post scores for 387 of these students, and also for 44 standard deductive-logic students and 23 straight critical-thinking-course students who also took their courses as freshmen (Hatcher, 1992, 1999; Sergeant, 1996).

EW 22. Twenty-four undergraduates in "Critical Thinking in Psychology" at a California State University unit, pre-tests compared with post-tests (Schultz, 1999).

EW 23. Fifty-two students who completed community college courses in U. S. History 1877 to the Present, twenty-nine of whom had Paul's model of critical thinking infused into their critical thinking instruction. Groups were compared on pre-post critical thinking differences, and on history content at the end of the courses (based on a document-based question from a disclosed College Board AP exam, and thirty-five multiple-choice questions from disclosed forms of *the College Board Achievement Test in American History and Social Studies* (Reed, 2001).

EW 24. Fifty-two first-year Philippine college students (age about 16 years) pre- and post-tested before and after a course in communication skills in which critical thinking skills were infused. (Lopez, 2004).

USER NORMS

User norms, which give group size, means and standard deviations of the particular groups studied, are not as helpful as would be norms based on stratified random sampling of populations of interest. However, such a sample is never really possible, and compromises and assumptions must be made. User norms are a rougher approximation than one usually finds, but at least the assumptions and compromises are more out in the open.

These user norms are separated into two tables, 1a and 1b (1a for groups without claimed prior critical thinking instruction, and 1b for groups with prior claimed or suspected critical thinking instruction).

In more detail: Table 1a provides pre-test results for experimental groups, pre- and/or post-test results for control groups, and results for other groups without claimed prior critical thinking instruction. Control groups with pre- and post-test results are treated as different groups and are thus included twice. Table 1b provides post-test results for experimental groups and other groups with claimed or possible prior critical thinking instruction. These data, where resulting from use in an experiment, are presented again later (reorganized) in Table 3, "Experimental Results".

The distinction between ‘no instruction’ (Table 1a) and ‘instruction’ (Table 1b) is sometimes difficult to make, partly because it is difficult to tell from the accounts provided exactly what happened. But another problem is in deciding whether certain descriptions, such as "instruction in technology" and "persuasive writing" indicate critical thinking instruction. The problem recurs in interpreting experimental studies, which I try to do later in connection with Table 3.

For convenience and a rough overview (a very rough overview, in case of only a few studies), unweighted means of means and standard deviations have been calculated, and are listed with total (not mean) number of students involved. These means are unweighted in order to emphasize the distribution, rather than the results for the larger groups.

Table 1a. Ennis-Weir User Norms for Students without Claimed Prior Critical Thinking Instruction

Notes:

1. The identification of a group (e.g., "EW12") refers to the listing of the group in the previous section, "The Nature Of The Groups Involved"

2. "EW" means Ennis-Weir. "E" means experimental group; "C" means control group. "SD" means standard deviation. "ct" means critical thinking. "N" means number of students. "NA" means not available.

Upper Elementary

Group	N	Mean	SD
EW12, prior to infusion of ct (pre-test)	27	6.0	4.2
EW12, prior to immersion approach to ct (pre-test)	22	3.4	4.8
Total N, unweighted means of means and SD's for these two groups	49	4.7	4.5

Middle School

Group	N	Mean	SD
EW5: above average 8th graders	93	2.1	NA
EW16: 7th & 8th graders — technology (pre-test)	38	6.7	3.5
Total N, unweighted means of means and SD's for these two groups	131	4.4	3.5

High School

Group	N	Mean	SD
EW6: C's in Jamaica (post-test)	32	1.4	6.3
EW8: C's in persuasive writing	82	6.6	4.8

(pre-test)			
EW8: C's in persuasive writing (post-test)	82	5.0	3.5
EW8: E's in persuasive writing, using Toulmin's model (pre-test)	66	7.8	5.5
EW10: Gifted C's in English (pre-test)	39	14.8	NA
EW10: Gifted C's in English (post-test)	39	17.0	NA
EW10: Gifted E's in philosophy (pre-test)	21	17.1	NA
EW11: (regular version of EW — without hints)	22	8.8	6.1
EW18: 10^{th}, 11^{th}, 12^{th} graders, test results to be factor analyzed	172	5.2	6.3
Total N, unweighted means of means and SD's for these nine groups	555	9.3	5.4

College

Group	N	Mean	SD
EW13: C's communications (post-test)	17	15.0	5.8
EW14: ed. psych. students tested to determine relationships	187	14.6	6.1
EW 19: one-semester argumentation course in English (pre-test)	84	13.0	7.3
EW21: freshmen, two semesters	977	7.5	5.3

in ct & writing (pre-test)			
EW22: ct in psychology (pre-test)	24	9.1	6.1
EW23: history C's (pre-test)	23	11.1	7.9
EW23: history C's (post-test)	23	8.5	8.3
EW23: history E's (pre-test)	29	11.9	8.6
Total N, unweighted means of means and SD's for these eight groups	1,364	11.3	6.9

Adult

Group	N	Mean	SD
EW7: C's: no lesson plan remodeling, no Paul frame (post-test)	16	6.0	7.0

Non-native Speakers of English, College

Group	N	Mean	SD
EW15: undergraduates in ct course (pre-test)	60	1.9	6.9
EW20: C's Asian Jr. College women in ESL/EFL* (post-test)	19	0.6	NA
EW24: first year college, ct infused in communications (pre-test)	52	1.3	2.2
Total N, unweighted means of	131	1.3	4.6

means and SD's for these three groups			

*English as a Second Language; English as a Foreign Language

Table 1b. Ennis-Weir User Norms for Students Who Have Received Some Claimed or Possible Critical Thinking Instruction

Notes:

1. The identification of a group (e.g., "EW12") refers to the listing of groups in the previous section, "The Nature Of The Groups Involved".

2. "EW" means Ennis-Weir. "E" means experimental group; "C" means control group. "SD" means standard deviation. "ct" means critical thinking. "N" means number of students. "NA" means not available.

Upper Elementary

Group	N	Mean	SD
EW12 after infusion of ct (post-test)	27	10.9	4.2
EW12 after immersion approach to ct (post-test)	22	9.1	4.6
Total N, unweighted means of means and SD's for these two groups	49	10.0	4.4

Middle School

Group	N	Mean	SD

EW3: Gifted 8th	28	18.6	5.9
EW9: a thinking-writing course - Gifted 7th& 8th	27	9.7	5.5
EW16: 7th & 8th graders — technology (post-test)	38	8.4	4.5
Total N, unweighted means of means and SD's for these three groups	93	12.2	5.3

High School

Group	N	Mean	SD
EW6: E's in Jamaica (post-test)	33	7.8	5.0
EW8: E's in persuasive writing (post-test)	66	5.7	4.6
EW10: Gifted E's in philosophy (post-test)	21	21.0	NA
Total N, unweighted means of means and SD's for these three groups	120	11.5	4.8

College

Group	N	Mean	SD

EW1a: after one semester of deductive logic (post test)	46	6.7	4.7
EW1b: after two semesters of writing & ct (post-test)	25	15.0	5.0
EW2: midway through a course in critical thinking/informal logic	27	23.8	4.0
EW17: after instruction involving identification of chemicals	29	7.0	3.5
EW13: E's in communications: ct & deconstruction (post-test)	16	20.8	3.1
Ew19: one-semester English argumentation course (post-test)	84	15.9	7.5
EW21: after two semesters ct & writing, freshmen (post-test)	977	12.8	5.7
EW 21 seniors who had ct & writing when freshmen (post test)	387	16.0	NA
EW22: ct in psych (post-test)	24	10.9	6.1
EW23: E's, ct in history (post-test)	29	15.2	8.8
Total N, unweighted means of means and SD's for these ten groups	1,644	14.4	5.4

Adult

Group	N	Mean	SD
EW7: remodeling lesson plans, no Paul frame (post-test)	15	8.7	10.1
EW7: remodeling lesson plans with Paul frame (post-test)	16	12.0	8.4

Total N, unweighted means of means and SD's for these two groups	31	10.4	9.3

Non-native Speakers of English, College

Group	N	Mean	SD
EW15: undergraduates in ct course (post-test)	60	11.4	9.0
EW20: E's, Asian women with ct infused in ESL/EFL* (post-test)	17	6.6	NA
EW24: first-year college, ct infused in communications (post-test)	52	8.2	6.0
Total N, unweighted means of means and SD's for these three groups	129	8.7	7.5

*English as a Second Language; English as a Foreign Language

"RELIABILITY" INDICES

In psychometrics, "reliability" means consistency. 'Situational validity' (Ennis, in process-B) means the extent to which a test successfully measures some specified thing in a situation. "Reliability" and validity are often confused because in everyday speech, 'reliability' implies getting it right, rather than just being consistent. A test can have very high "reliability" and actually totally fail to assess what it is purported to assess.

The problem is exacerbated, mostly for multiple-choice tests, when, as commonly calculated, "reliability" is a measure of internal consistency among items, rather than consistency from one time or grader to another. A test that tests for a multidimensional concept (like *critical thinking*) is thus penalized. For the Ennis-Weir test, an internal

consistency index like a Kuder-Richardson or the Cronbach alpha is inappropriate because of the multidimensionality of the concept, *critical thinking*, the paucity of items, and the dependence of the ninth item on the others. However, the Cronbach alpha has been calculated by two researchers (see EW 14 and EW 18), and is not as low as one might expect from the first and second inappropriateness problems just mentioned. See Ennis (2000) for a longer discussion of "reliability".

The Ennis-Weir test, however, has a record of good inter-rater "reliabilities" for high school and college students, and for gifted younger students.

Table 2. Obtained "Reliability" Indices

Group	Level	N	"Reliability"	Type
EW 2	College	27	.86	Inter-rater
EW 3	8th grade gifted	28	.82	Inter-rater
EW 5	8th grade "above average"	93	.97	Inter-rater
EW 12	5th and 6th grade	49	.58 pre; .61 post	Inter-rater*
EW 14	College	187	.59**	Cronbach alpha
EW 15	College, Eng. Second Language	60	.91 pre; .92 post	Inter-rater
EW 17	College	29	.92	Same rater, 4 months stability
EW 18	High school	172	.72**	Cronbach alpha
EW 19	College	25	.90	Inter-rater
EW 20	College, Eng Second	36	.72	Inter-rater

	Language			
EW 21	College	101	.93	Inter-rater
EW 22	College	48	.74	Inter-rater
EW 23	College	52	.98 pre, .99 post	Inter-rater

*The author remarked that the graders were very generous, which fact, together with the index obtained, suggests that these students were too young for this test, given that they were not gifted.

**As mentioned above, this method is not well suited to an essay test, but the result is interesting.

SITUATIONAL VALIDITY

In this supplement, I employ an approach to test validity that I have developed recently and that responds to the apparent impasse between the original concept of test validity and the view that validity is not a property of tests at all.

Test validity, we were once told, is the extent to which a test assesses what it is supposed to assess. But this approach founders on the fact that test results depend on things other than what a test is supposed to assess, including whether the test is in the first language of the students taking it, and whether it assumes some cultural knowledge unfamiliar to the students.

A psychometric-leadership response to this situation is to deem that validity is not a property of tests, but rather of inferences drawn from, and/or, interpretations of, test scores (Messick, 1989a, 1989b; Joint Committee…, 1999; Frisbe, 2005). But which inferences, which interpretations? Furthermore, the response seems to neglect our interest in the test itself.

So I suggest (Ennis, in process-B) that we situationalize test validity, that is, focus on the extent to which *in the situation* the test assesses what it is supposed to assess. This works conceptually for test users because that is what they are really interested in. But for someone writing about the test in general, as I am doing in this supplement, some generalization is necessary because it is to be read by people in different situations. So I talk here about standard-condition situational validity, which is the extent to which the test in standard situations assesses what it is supposed to assess; and I realize that there will be situations where it will not do this very well, such as when the test is taken by non-native

speakers of English. There are some results reported in this supplement for non-native speakers, and the results are just not comparable to those for native speakers of English.

There is much more to be said about this issue, but in any case, one must be especially careful about comparing groups, using the test in experiments, and deciding about students' critical thinking prowess -- when situations are not standard.

EXPERIMENTAL RESULTS

There are not enough focused and well-controlled studies to give definitive empirical answers to these three interrelated significant questions:

1) Are attempts to teach critical thinking successful?

2) Is the Ennis-Weir test a situationally-valid test in standard conditions?

3) Does the Ennis-Weir test assess dispositions as well as abilities?

But the results are generally consistent with an affirmative answer to each. To elaborate on each of these questions, one at a time:

Question 1, Teaching Critical Thinking

First, a qualification: Realistic studies of the question about teaching of critical thinking tend to have problems because they are done in real educational settings where many other factors

Table 3. Experimental Results

Notes:

1. All of these data have appeared in Table 1a or 1b, but are reorganized here for purposes of understanding experimental results.

2. "Sig." means statistically significant. "E" means experimental group. "C" means control group. "G" means gifted. "ct" means critical thinking. "SD" means standard deviation. "N" means number of students. "NA" means not available

Upper Elementary

		Pre				**Post**		**Comment**
Group								

	N	Mean	SD		N	Mean	SD	
EW12 (infusion)	27	6.0	4.2		27	10.9	4.2	Sig. improvement for each;
EW12 (immersion)	22	3.4	4.8		22	9.1	4.6	no sig. diff. reported between
								improvements

Middle School

		Pre				**Post**		**Comment**
Group	N	Mean	SD		N	Mean	SD	
EW16 (7^{th} to 8^{th},	38	6.7	3.5		38	8.4	4.5	Sig. improvement
Technology)								

High School

		Pre				**Post**		**Comment**
Group	N	Mean	SD		N	Mean	SD	
EW6 C's					32	1.4	6.3	E's sig. better after ct instruction,
EW6 E's					33	7.8	5.0	which closely followed EW test

								in form and content
EW8 C's	82	6.6	4.8		82	5.0	3.5	Persuasive writing: both groups
EW8 E's	66	7.8	5.5		66	5.7	4.6	worsened
EW10 C's G Eng.	39	14.8	NA		39	17.0	NA	Both E's and C's improved, no
EW10 E's G Phil	21	17.1	NA		21	21.0	NA	SD's supplied

College

		Pre				**Post**		**Comment**
Group	N	Mean	SD		N	Mean	SD	
EW1 ded. logic					46	6.7	4.7	Sig superiority of writing and ct
EW1 two semes-, ters,writing & ct					25	15.0	5.0	over ded. logic (+8.3); Cohen's d= 1.7
EW13 C's					17	15.0	5.8	Sig diff (+5.8), deconstructive

								ct E's over standard
EW13 E's					16	20.8	3.1	communications C's
EW19 English argumentation	84	13.0	7.3		84	15.9	7.5	Sig. improvement (+2.9); Cohen's d*= 0.4
EW 21 freshmen, writing & ct	977	7.5	5.3		977	12.8	5.7	Sig improvement (+5.3); Cohen's d*=0.9
EW 21 same course as above at freshman level, tested again as seniors	387	7.9	NA		387	16.0	NA	Improvement (+8.1) reported sig.; no SD's provided
EW 21, standard ct.course as fresh- men; tested again	23	12.1	NA		23	13.7	NA	Improvement (+1.6); no SD's provided

as seniors								
EW 21, standard ded. logic as freshmen; tested again as seniors	44	11.2	NA		44	9.5	NA	Decrease (-1.7); no SD's reported
EW22 ct in psych	24	9.1	6.1		24	10.9	6.1	Sig. improvement (+1.8)
EW23 C's (hist.)	23	11.1	7.9		23	8.5	8.3	Diff of -2.6; a decrease; puzzling
EW23 E's (ct in hist.)	29	11.9	8.6		29	15.2	8.8	Diff of +3.3; sig diff between diffs. of E's and C's; "effect size" =.83; E's and C's equal on historical content.

*Cohen's d is an indicator of size of effect: ratio of difference between means to SD (Cohen, 1992).

Adult

		Pre				**Post**		**Comment**
Group	N	Mean	SD		N	Mean	SD	
EW7 (base line)					16	6	7.0	No sig.
EW7 (no frame)					15	8.7	10.1	diffs
EW7 (w/frame)					16	12.0	8.4	claimed

Non-Native Speakers of English

		Pre				**Post**		**Comment**
Group	N	Mean	SD		N	Mean	SD	
EW 15	60	1.9	6.9		60	11.4	9.0	Sig. improvement (+9.5)
EW20 C's					19	0.6	NA	Sig. diff (+6.0) reported, though
EW20 E's					17	6.6	NA	SD's were not reported
EW 24	52	1.3	2.2		52	8.2	6.0	Sig improvement (6.9)

Internal Analysis of the Ennis-Weir Test: Critical thinking Dispositions

	Group	N	Result	Comment
Internal: CT dimensions: dispositions and abilities. Experimental variable: hints	EW 11	64	Means went from 8.8 (6.0, 22) to 9.4 (6.4, 21) to 14.5 (6.8, 21), (SD's and N's in parentheses.)	In two of three administrations, hints to make up for absence of dispositions were accompanied by higher scores. The first mean was with no hints, the second was with general hints, and the last was with specific hints. SD and N are in parentheses.

impinge (such as scheduling, student interests, pressures to do well on certain achievement tests that do not assess critical thinking, difficulty of obtaining control groups, difficulty of random assignment to experimental and control groups when control groups are possible, lack of motivation among, and hence cooperation by, control-group students, different conceptions of critical thinking, difficulty of generalizing from a set of unique cases, etc.).

In spite of these difficulties, in most of the studies in which an attempt was made to teach critical thinking, there were significant differences reported favoring the taught group,

suggesting a positive answer to Question 1. In elaborating on this loose generalization, I shall refer to the findings by group identification in Table 3, "Experimental Results", but will not treat every study in detail. Please refer to the original studies for more information.

This elaboration assumes that persuasive writing and deductive logic (as taught in symbolic logic courses) are not critical thinking (Ennis, 1962, 1981, 1987, 1991, 2002). Persuasive writing (EW8) is different from critical thinking in that its goal is persuasion, while the goal of critical thinking is to try to get it right. The control group in EW8 was taught persuasive writing, and its mean score did not improve, as one would expect, given the basic difference between the goals of trying to persuade and trying to get it right.

An added wrinkle, which I would also expect, is that the experimental group in EW8 (which was also taught persuasive writing, but in accord with the Toulmin model, 1964, pp. 97-107), also did not improve. This result also is to be expected because the Toulmin model gives only the structure of an argument (conclusion, reasons(s), warrant, qualifiers, rebuttal, etc.), but no advice about how to try to get right a judgment, belief, or decision. Incidentally, Hitchcock (2005) has tried to fill this gap in the Toulmin model, but the original model did not have the needed advice. So the lack of improvement in the EW8 experimental group, as well as the EW8 control group, is to be expected.

Deductive logic (taught in the logic classes in EW1 and EW 21) as generally taught these days is heavily symbolic with elaborate systems that go far beyond what is needed from deductive logic in critical thinking and introduce features (like the paradoxes of material implication) that are counterintuitive and confusing to students (Ennis, 1981; Lewis, 1912; Strawson, 1952). So I would not expect a deductive logic class of that sort to help students become better critical thinkers.

In sum, neither persuasive writing (with or without the Toulmin model) nor deductive logic was found to improve critical thinking (EW8, EW1 and EW21), a result that one might expect.

In the following experimental studies, the teaching was claimed to be successful: (EW1, EW12, EW6, EW 13, EW19, EW21, EW22, EW23, EW24, EW7, EW15, EW20, and EW24. In two of these studies (EW1 and EW21), instruction for the experimental groups combined critical thinking with writing. Their success is not surprising because they seemed to be actually teaching critical thinking and because the Ennis-Weir test requires writing. For various similar types of consideration, the results of the other apparently successful attempts are not surprising. Basically the general principle is that students tend to learn what they are taught if the conditions are favorable. To be more precise at this point is not warranted because of the variety of specific situations in these studies. But these results fit in with the finding reported in Ennis, Millman & Tomko (2005) that explicit attention to the principles of critical thinking is generally helpful.

Some of these studies did not have control groups and were pre-post-test studies only, a problem for definitive proof because there are other explanations possible. But still there is a fairly strong trend here, which fits in with the truism that generally a good way to teach something is to teach it deliberately.

Consistent with the findings with the Cornell critical thinking tests (Ennis, Millman & Tomko, 2005), the one study with the Ennis-Weir that checked the impact of the infusion of critical thinking on content learning (EW23) found that experimental and control students at the college level did equally well on history content.

The one study (EW12) that compared the infusion and immersion approaches to teaching critical thinking (see Ennis, 1989) found a significant improvement for each, but no significant difference between them. However,the study was done with fifth graders and had fairly low inter-rater consistency (.58 and .61), so repetition with older groups is needed.

Question 2. The Standard-Condition Situational Validity of the Ennis-Weir Test

The strongest evidence for standard-condition situational validity is content-related and is presented in the original test/manual. One important aspect is that the test is based on a defensible and well-elaborated conception of critical thinking (Ennis, 1962, 1981, 1987, 1991, 1996, 2002), which was deemed "the prevailing view of critical thinking" by critic John McPeck (1981, Ch. 3).

One significant empirical indicator of situational validity (Messick, 1989, p. 6; Ennis, in process-A) is a test's showing improvement in something that has been taught in an experiment — and not showing improvement on something that was not taught (given appropriate assumptions). On this criterion, the Ennis-Weir test does very well, as can be seen from the discussion of Question 1.

The basic idea here is that the test's being situationally valid (a hypothesis), together with other factors, such as there having been successful teaching of critical thinking, explains the favorable results on the test. Comparably, the hypothesis, together with the assumption that critical thinking was not taught (and, with appropriate other assumptions) would explain lack of improvement on the test. Explanatory power gives support to the hypothesis (Harman, 1973).

A second empirical indicator is the test results' making good sense. As can be seen in this and the next section, "Relationships", the test does well on this criterion also.

Question 3. Does the Ennis-Weir test assess critical thinking dispositions, at least in part?

In an ingenious study with Group EW11, Norris & Hollett (1992; also Norris, 2003) gave three different versions of the Ennis-Weir test. One was the standard version. The second was the standard version plus general hints in the initial directions, such as, "consider the whole situation" and "think of alternatives". The third provided a specific hint inserted after each paragraph, such as, "think about other explanations for the results" after Paragraph 6, which deals with the police chief's traffic study. As I interpret the study, Norris & Hollett's thinking was that if the student does not have the disposition to think about other explanations for the results, but receives that hint after Paragraph 6, the lack of the disposition would be compensated by the hint, and the student would then be likely to do better on the item than the student would do without the hint. A student whose score improves with the specific hint would then be less likely to have the disposition than one whose score was not improved by the specific hint. This makes the score of the original test dependent on the student's having the disposition in the first place, and thus in part a test of critical thinking dispositions.

The results of this study, as presented at the end of Table 3 show a mean total score 8.8 on the original test, 9.4 with the general hints added to the initial directions, and 14.5 with a specific hint inserted after the paragraph. This study needs replication and variation, but the results suggest that the Ennis-Weir test is at least in part a test of critical thinking dispositions. The study also supports the viability of the distinction between critical thinking dispositions and abilities. Further evidence to support the viability of the distinction is presented later in the "Personality" part of the "Relationships" section, and in Taube (1993), reporting a factor analysis of the test with Group EW14, also presented in the "Relationships" section.

RELATIONSHIPS

The relationships between critical thinking and other variables and factors that were found in these studies contribute to our general knowledge about the world of human beings, culture, and education. To the extent that they contribute and to the extent that they make sense as parts of broader theories and views of human beings, they also contribute to the case for the standard-condition situational validity of the Ennis-Weir test (Messick, 1989, p. 6; Ennis, in process-A). Basically the goal in a case for standard-condition situational validity is that the test results make sense; more specifically, that the test results are well-explained by the hypothesis that the test is situationally valid in standard conditions, and better explained by this hypothesis than by alternative hypotheses.

Types of relationships examined include Ennis-Weir-Critical-Thinking-Essay-Test correlations with other putative measures of critical thinking, academic variables, personality variables, and gender. The findings are not extensive, but they do make sense and are presented in "Table 4, Relationships".

Other Putative Measures of Critical Thinking

A low to moderate relationship (.37, .28) was found with the *Watson-Glaser Critical Thinking Appraisal* (EW 14 and EW18). This is to be expected because although both tests claim to assess critical thinking, 1) the conception of critical thinking are different; 2) while one is a multiple-choice test, the other calls for students to construct their own answers; and 3) the Watson-Glaser test still has some of the difficulties that it had in the middle of the last century (Ennis, 1958). The Ennis-Weir test also had low-moderate relationships (EW18) with two multiple-choice tests, each of which tested for only one aspect of critical thinking, making observations, and doing best-explanation reasoning. Unfortunately, I was unable to find correlations with either the complete Level X or Level Z of the Cornell critical thinking tests (Ennis & Millman, 1985a&b), or with any other leading tests (other than the Watson-Glaser). Such correlations are hereby solicited.

Academic Achievement/Aptitude

For students with English as a first language, low/moderate correlations with SAT verbal (EW14: .40), SAT Quantitative (EW14: .28) were obtained. These results seem reasonable in the light of the fact that there was no special attention to critical thinking in the SAT at that time.

Table 4. Relationships

Critical Thinking

Variable	**Group**	**N**	**Relationship**	Comment
Watson-Glaser CT Appraisal	EW14	187	.37	Not surprising, given the differences between the tests
Watson-Glaser CT Appraisal	EW 18	172	.28	Not surprising, given the

				differences between the tests.
Section I, Cornell CT Test,Level X (best-explanation reasoning). (Ennis & Millman, 1985)	EW18	172	.32	Only the best-explanation section of Level X (23 multiple-choice items) was used
Test on Appraising Observations (constructed-response version), (Norris, 1986)	EW18	172	.25	Observation appraisal is one of a number of aspects of critical thinking

Personality

Variable	**Group**	**N**	**Relationship**	Comment
AT20 (ambiguity tolerance)	EW14	187	.33	
NCS (need for cognition)	EW 14	187	.24	
CLEV (Perry's educational values)	EW 14	187	.35	

Academic

Variable	**Group**	**N**	**Relationship**	Comment
SAT* Verbal	EW14	155	.40	
SAT Quantitative	EW14	155	.28	
Grade Point Average	EW14	171	.28	With the Cornell tests (Ennis, Millman & Tomko (2005), the relationship with GPA varied considerably from one institution to another, as is to be expected.
SAT Verbal	EW15	60	.34	EW* Pre-test
SAT Verbal	EW15	60	.59	EW Post-test
TOEFL*	EW15	60	.35	EW Pre-test
TOEFL	EW15	60	.48	EW Post-test
ACT	EW15	60	.25	EW Pre-test
ACT	EW 15	60	.66	EW Post-test

TWE*	EW15	60	-.56	EW Pre-test
TWE	EW15	60	-.07	EW Post-test
SPM*	EW15	60	.41	EW Pre-test
SPM	EW 15	60	.35	EW Post-test

* SAT = Scholastic Achievement Test. EW = The Ennis-Weir Critical Thinking Essay Test. TOEFL = Test of English as a Foreign Language. TWE = Test of Written English (see text). SPM is the composite score for the national-level Maylaysian Certificate of Education.

Other

Variable	**Group**	**N**	**Relationship**	Comment
Internal: CT dimensions: dispositions and abilities	EW 14	187	Distinction supported	Factor analysis used
Gender	EW5	93	Females better	Sig. better on pre-test; "generally outperformed males" (8th grade)
EW Pre-post relationship, given	EW21	800	.55	This is not "reliability" because

intervening ct instruction				of the intervening instruction
EW Pre-post relationship, given intervening ct instruction	EW19	84	.79	This is not "reliability" because of the intervening instruction
EW Pre-post relationship, given intervening ct instruction	EW22	24	.82	This is not "reliability" because of the intervening instruction

The same group provided a correlation with college grade point average of .28, which is in about the middle of the wide range of correlations with grade point average (from .63 to -.02) found

with the Cornell critical thinking tests (Ennis, Millman, & Tomko, 2005). This wide range is consistent with the wide range in emphases on critical thinking that we find in our education system. So the Ennis-Weir results again are not surprising.

An interesting series of results for non-native speakers of English was obtained with EW15, 60 Maylaysian undergraduates at a large Midwestern university taking courses in critical thinking and English speaking and writing, among other things. Both the SAT Verbal and the ACT test correlated low/ moderate with the Ennis-Weir pre-test (.34 and .25) which is consistent with the results with EW14. However, these tests were much better at predicting the Ennis-Weir post-test results with correlations of .59 and .66. A similar, though not so pronounced result was found with TOEFL (Test of English as a Foreign Language): .35 going up to .48. With TWE (Test of Written English) which only calls for narrative writing, the movement was in the same direction but it went from a negative .56 to a negative .07! this negativity needs further exploration, but suggests that critical thinking might not be, or might be negatively, related to narrative ability.

The correlation with the Maylaysian Certificate of Education score, on the other hand, reduced from .41 to .35, Ennis-Weir pre-test to post-test, which is consistent with the proposition that this instrument depends more heavily on memorization than the others

and is not as good a predictor of later critical thinking success -- but still is better than the TWE.

Personality

Low/moderate correlations with the personality variables assessed in EW14 are what one would expect and desire. These variables are tolerance of ambiguity (AT20, r = .33); need for cognition, which is basically the tendency for an individual to engage in and enjoy thinking (NCS, r = .24); and Perry's educational values (CLEV, r = .35), which go up the scale of intellectual development from absolute polarization of everything to tolerance of ambiguity.

Critical Thinking Dispositions

These EW14 personality correlations, together with the fact that these personality variables are actually critical thinking dispositions, support the proposition that the Ennis-Weir test is in part a test of critical thinking dispositions. A confirmatory factor analysis done by the EW14 author (Taube, 1993), reported earlier, is consistent with this proposition. Further confirmation is found in the study by Norris & Hollett (1992) with EW11, in which presence and specificity of hints accompany the Ennis-Weir test were varied, reported earlier under "Experimental Results".

It is to be expected that the Ennis-Weir test would test in part for critical thinking dispositions, given its nature. Paragraph 6, for example, expects students to exercise the disposition to be alert for alternatives in order to do well in responding to it.

Gender

The only Ennis-Weir study I found (EW 5) that compared genders was consistent with the rather consistent results of the studies with the Cornell tests (Ennis, Millman, & Tomko (2005, pp. 26, 36), in which at the lower educational levels females did better than males, but that at the higher levels they were about the same. In EW5, eighth grade girls "generally outperformed males").

Pre-Post-Test Relations with Intervening Instruction

Correlations between pre- and post-tests with intervening instruction cannot be considered "reliability" indices. However the correlations are interesting because they show that the Ennis-Weir as a pre-test is a fairly good predictor of post-test scores and thus might be used as a co-variate. Correlations of .55, .79, and .82 were found in EW21, EW19, and EW22.

SUMMARY

In this supplement can be found a set of user norms for students without previous claimed instruction in critical thinking as well as a set for students having experienced claimed or possible instruction in critical thinking. Also contained are a set of "reliability" indices that are quite high for an essay test (the word ‘reliability’ being in double quotes to remind the unwary that psychometric reliability is consistency and could be quite high with a totally invalid test).

Empirical support for a standard-condition situational validity claim for the test is to be found in results of experiments as well as some identified relationships with other variables: 1) Experimental results support the truism that students learn what we try to teach them if conditions are satisfactory, but they also give support to a standard-condition situational validity claim for the Ennis-Weir test because they are partially explained by that claim. That is, the standard-condition situational validity of the test together with the truism explains why experimental students generally did better on the test. 2) Empirically-ascertained relationships (generally correlations) are what we might expect with certain critical thinking tests, academic prowess, personality, and gender, but more studies are needed to confirm these results. Findings with the test that accord with what we would expect are also supportive of the case for standard-condition situational validity.

The findings in two studies of a dispositional component of the Ennis-Weir test also support the situational validity of the test because we would expect the results of an open-ended essay test like this to depend in part on the presence of critical thinking dispositions.

Strong content-related evidence of standard-condition situational validity is provided in the original test/manual booklet (Ennis & Weir, 1985). Basically, the argument is that the test covers reasonably well the basic components of an acceptable conception of critical thinking and does so realistically.

Findings from the test are in accord with other research in the area (especially that reported in the latest Cornell critical thinking tests manual (Ennis, Millman, & Tomko, 2005): critical thinking instruction is more likely to be effective if it is done directly with explicit attention to critical thinking; girls tend to do better at critical thinking at earlier ages but boys tend to catch up by the college years; and subject-matter acquisition is not sacrificed by the infusion of critical thinking into a curriculum. However, many more studies are needed if these findings are to be regarded as established — or refuted -- and for us to make finer distinctions.

In brief, the Ennis-Weir test has strong content-related support and moderate empirical support for a claim for its situational validity under standard conditions; its "reliability" is strong; and the tentative findings from its use are interesting. But more studies are needed.

REFERENCES

Chennault, Anita (1989). Enhancing critical thinking skills in gifted elementary school students. Unpublished doctoral dissertation, Mississippi State University. (EW 4)

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 1, 155-159.

Colledge, Deborah G. (1993). An experimental study comparing embedding and immersion approaches to instructing critical thinking to 5th grade students. Unpublished doctoral dissertation, Pennsylvania State University. (EW 12)

Davidson, Bruce W. and Dunham, Rodney A. (1997). Assessing EFL student progress in critical thinking with *The Ennis-Weir critical thinking essay test*. *JALT Journal, 19*, 1 (May), 43-57. (EW 20)

Ennis, R.H. (1958). An appraisal of the *Watson-Glaser critical thinking appraisal*. *Journal of Educational Research, 52*, 155-158.

Ennis, R.H. (1962). A concept of critical thinking. Harvard Educational Review, 32, 81-111.

Ennis, R.H. (1981). A conception of deductive logic competence. Teaching Philosophy, 4, 337-385.

Ennis, R.H. (1987). A taxonomy of critical thinking dispositions and abilities. In J. Baron & R. Sternberg (Eds.), Teaching thinking skills: Theory and practice. New York: W.H. Freeman. Pp. 9-26.

Ennis, R.H. (1989). Critical thinking and subject specificity: Clarification and needed research. Educational Researcher, 18 (3), 4-10.

Ennis, R.H. (1991). Critical thinking: A streamlined conception. Teaching Philosophy, 14 (1), 5-25.

Ennis, R. H. (2000). Test reliability: A practical exemplification of ordinary language philosophy. Philosophy of education 1999.Champaign, IL: Philosophy of Education Society.

Ennis, R. H. (2002). Goals for a critical thinking curriculum and its assessment. In Arthur L. Costa (Ed.), Developing minds (3rd Edition). Alexandria, VA: ASCD. Pp. 44-46.

Ennis, R. H. (in process-A). Investigating and assessing multiple-choice critical thinking tests. In Sobocan, Jan & Groarke, Leo (Eds.), probable title: *Teaching and testing: Critical thinking in today's schools and universities*.

Ennis, R. H. (in process-B). Situational test validity.

Ennis, R. H. & Millman, J. (2005a). Cornell Critical Thinking Test, Level X. Pacific Grove, CA: Midwest Publications.

Ennis, R. H. & Millman, J. (2005b). Cornell Critical Thinking Test, Level Z. Pacific Grove, CA: Midwest Publications.

Ennis, R. H., Millman, J. & Tomko, T. N. (2005). *Cornell critical thinking tests Level X & Level Z Manual*, (Fourth Edition). Seaside, CA: The Critical Thinking Co.

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

Frisbie, D.A. (2005). Measurement 101: Some fundamentals revisited. *Educational Measurement: Issues and Practice, 24* (3), Fall, 21-28.

Goldberg, Mary Lou (1991). A study of critical thinking competencies in above-average eighth-grade students. Unpublished doctoral dissertation, Temple University. (EW 5)

Harman, Gilbert (1973). *Thought*. Princeton, NJ: Princeton University Press.

Hatcher, Don (1999). Why we should combine critical thinking and written instruction. *Informal Logic, 19*, 2 & 3 (Summer & Autumn), 171-183. (EW 21).

Hatcher, Don (1992). The new synthesis. *CT News, 10*, 3&4 (Feb, Mar, Apr, May), 1-5. (EW 21)

Hitchcock, David (2005). Good reasoning on the Toulmin model. In Hitchcock, David (Ed.), *The uses of argument: Proceedings of a conference at McMaster University*.Hamilton, Ontario: Ontario Society for the Study of Argumentation.

Jacks, Mary Jane (1994). An evaluation of a model technology program intended to develop critical thinking abilities of junior high school students. Unpublished doctoral dissertation, University of Southern California. (EW 16)

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

Koehler, Caril F. (1993). The effectiveness of deconstructing teaching strategies on students' critical thinking skills. Unpublished doctoral dissertation, University of Missouri-Kansas City. (EW 13)

Lewis, Charles I. (1912). Implications and the algebra of logic. *Mind* (October), 522-531.

Lopez, Marcus (2004). Development and validation of critical thinking infusion lessons in communication skills for freshman college students. Unpublished doctoral dissertation, Philippine Normal University. (EW 24).

Messick, S. (1989a). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed.) (13-103). New York: Macmillan.

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

Moore, Rashid A. (1995). The relationship between critical thinking, global English language proficiency, writing, and academic development for 60 Maylaysian second language learners. Unpublished doctoral dissertation, Indiana University. (EW15)

Moreyra, Alicia (1991). The role of thinking frames in developing teachers' critical thinking skills and dispositions. Unpublished doctoral dissertation, University of Miami. (EW 7)

Morgan, Wayne R. (1995). Assessment of critical thinking strategies utilized by community college students. Unpublished doctoral dissertation, Kansas State University. (EW 17)

Norris, Stephen P. (1995). Format effects on critical thinking test performance. *The Alberta Journal of Educational Research, 41,* 4 (December), 378-406. (EW 18).

Norris, Stephen P. (2003). The meaning of critical thinking test performance: The effects of abilities and dispositions on scores. In Fasko, Daniel, Jr. (ed.), *Critical thinking and reasoning: Current research, theory, and practice*. Cresskill, NJ: Hampton Press, Inc. Pp. 315-329. (EW11)

Norris, Stephen P. and Hollett, Ann (1992). Two issues concerning the validity of multiple-choice and constructed-response critical thinking tests: Their equivalence and

dependence upon critical thinking dispositions. A paper prepared for the Fifth International Conference on Thinking, Queensland, Australia, July 6-10. (EW 11).

Rapaport, William (1979). Personal communication. (EW1).

Reed, Jennifer H. and Kromrey, Jeffrey D. (2001). Teaching critical thinking in a community college history course: Empirical evidence from infusing Paul's model. *College Student Journal, 35* (June, 2001), 201-215. (EW 23)

Sergent, Susan D. (1997). Personal communication from Don Hatcher's statistician at the time. (EW 21)

Schultz, Wes (1999). Personal communication. (EW 22)

Sirbratthie, Norma (1991). The effect of a teaching programme in critical thinking. Unpublished masters thesis, The University of the West Indies. (EW 6)

Strawson, Peter F. (1952. *Inbtroduction to logical theory*. London: Methuen & Co., Ltd.

Taube, Kurt T. (1993). Critical thinking ability and disposition as factors of performance on a written critical thinking test. Unpublished doctoral dissertation, Purdue University.

Toulmin, Stephen (1964).*The uses of argument*. Cambridge, UK: Cambridge University Press.

Waldron, James Michael (1992). The effects of two instructional strategies for critical thinking-writing instruction on high ability middle school students. Unpublished doctoral dissertation, University of California, Riverside. (EW 9)

Wallace, Sally Pritchard (1992). A study of argumentative/persuasive writing related to a model of critical thinking in grades nine and eleven. Unpublished doctoral dissertation, University of Southern California. (EW 8)

Weiner, Linda (1996). Three critical thinking test instruments and the assessment of argumentation skills. Unpublished doctoral dissertation, University of Akron. (EW 19)

Yarbrough, Douglas B. (1992). Effects of an introduction to philosophy course on self, esteem, epistemology, and critical thinking ability of high ability adolescents. Unpublished doctoral dissertation, Walden University. (EW 10)