VII.2&3, Spring&Fall 1985
Informal Logic

Teaching Critical Thinking at the University Level:

A Review of Some Empirical Evidence*

LEONARD E. GIBBS *University of Wisconsin—Eau Claire*

This review was conducted specifically to help us plan a critical thinking program for faculty at the University of Wisconsin—Eau Claire, and to evaluate its effects on faculty and students in their classrooms. It seemed appropriate to first weigh evidence before beginning such a program. Thus, the questions listed in the abstract concern measures of critical thinking, effectiveness of conventional curricula, effectiveness of curricula designed specifically to teach critical thinking, and factors associated with successful learning by participants. Because we thought empirical studies would provide the clearest answers to our questions, studies of critical thinking in universities are the only evidence included.

Some clarifications may be helpful. Some otherwise excellent studies were excluded because they evaluated critical thinking at the pre-university level (Noyce, 1970; Smith & Tyler, 1942). The review begins with two tables. The first summarizes methodological features of studies reviewed; the second summarizes findings and measures used to quantify those findings. Readers who want a detailed overview of each study and a feature-by-feature comparison of it with other studies, may find the tables and their explanations helpful. Readers who want a quick overview of the evidence may want to skip the tables and read the discussion. Discussion hits high points in the tables and follows the sequence of questions posed in the abstract.

Criteria for Inclusion

When choosing evidence, I intended to cast a net with a wide enough aperture to catch the best empirical evidence but not so wide that it caught a confusing mixture of weak and strong evidence. Ideally, the best criteria for inclusion would be: random selection of subjects, measures of proven validity and reliability, random assignment of subjects to alternate programs for teaching critical thinking or to a control, specific hypotheses tested by appropriate inferential statistics, and sufficient follow-up to measure strength of effect over time. These criteria were too rigorous. The first sweep of the net caught nothing.

Nine studies did meet the following criteria: their authors studied effects of university level programs for teaching critical thinking; authors stated specifically that they were evaluating critical thinking; they used at least one measure of critical thinking to evaluate effects of teaching; they made some comparison (either pretest against posttest or across groups), and authors used descriptive or inferential statistics.

Because none had sufficient control over their experiment to randomly assign subjects to experimental conditions, designs summarized here are all quasi-experimental (Campbell and Stanley, 1963). Though there is no question that random assignment to alternate programs would enable more powerful causal inferences, random assignment is not the all inclusive facilitator of high quality research (Cook and Campbell, 1979). Thus, studies reported here, from various contexts, often using different measures, can still provide tentative answers to our

questions.

We located studies by asking experts for references, by reading reviews on the subject (Norris, 1985; Baker, 1979), and by searching DIALOG's ERIC and other files for the intersect "Critical Thinking" and "Higher Education." The review reflects monthly reviews of ERIC files through February, 1986.

Explanations of Tables

Table 1 shows how well each study meets several criteria for methodological precision. It describes each study's merits and allows a quick comparison across studies by criterion. The first column identifies each study by author and year. The second column contains the location of the study, where possible, and identifies the type of class or setting for subjects. Column three describes study design according to Cook and Campbell's (1979) terminology. The fourth and fifth columns give the number of subjects pretested and the number posttested, thus providing a quick reference to the number of subjects involved in the experiment and any subject attrition. The symbols Rs and Ra in columns six and seven respectively, denote whether subjects were randomly selected for inclusion in the study, or were randomly assigned to alternate treatments or to control. A slash (/) through these symbols means randomization criteria were not met. Column 8 lists the period of follow-up, or the interval between pretest and posttest, if such a design is used.

Column 9 lists the Credibility Index (CI) for each study. This index is based on a Quality of Study Rating Form that lists nine criteria for a good evaluation study and accompanying instructions for identifying and weighting those evaluation criteria (Gibbs, 1985). Thirty-nine raters have used the form to rate two studies agreeing an average of 95% and 93% with keyed criteria. Stronger randomized trials generally score above 70 points on the form. CI is computed by adding weights for the following criteria: random selection of subjects (10 points), random assignment (20 points), nontreated control or comparison group (10 points), number of subjects in the largest treatment group exceeding twenty (10 points), a check of validity by correlating the principal outcome measure with another similar measure (16 points), a reliability coefficient for the principal measure of critical thinking (15 points), a reliability coefficient of at least .70 or 70% agreement between raters (9 points), follow-up longer than six months (4 points), and using an inferential statistic to test comparisons for statistical significance (6 points). The CI can range from zero, in a study where none of the criteria are met, to one hundred, where all criteria are met.

Table 2 lists measures of critical thinking, criteria for evaluating measures, and summaries of study results. The first column identifies each study by author and year. The second describes the location and type of university class providing subjects. Columns 3 through 5 list respectively, the name of the measure or measures used to quantify critical thinking, the reliability coefficient or percent of inter-rater agreement for each measure, and information relevant to validity. Column 6 lists principal hypotheses; these may be explicitly stated by the author or implicit. Column 7 lists the statistical test and "p" (significance) level related to each hypothesis. (Here "p" level generally means the probability that a given result could be found due to chance alone; so the smaller the "p" level the greater our confidence in difference reported.) Column 8 lists the strength of treatment effect (SE) in standard deviation units (Glass, 1972; Hedges, 1984). This index is usually the mean of the experimental group minus the mean of the control group, all divided by the standard deviation of the control group, or the difference between treatments all divided by a pooled estimate of their

standard deviation. Especially pertinent comments are in column 9.

Findings

Which kinds of instruments have been used most frequently by evaluators to measure university level critical thinking? Column 3 of Table 2 shows that the Watson-Glaser Critical Thinking Appraisal, a test whose forms A and B were copyrighted in 1951, is most popular: three authors used it. The eighty-item Watson-Glaser is a multiple choice test of ability to discriminate among degrees of support for inferences, recognition of unstated assumptions, ability to make logical deductions, interpretation of evidence to see if generalizations or conclusions are warranted, and ability to judge the relevance of arguments to particular questions (Watson and Glaser, 1980). Two studies used a procedure for grading essay tests developed by Browne, Haas and Keeley (1978). Their rubric scores the following elements in student essays: identifying a controversy and conclusions regarding that controversy, identifying major arguments, identifying and analyzing implicit premises, recognizing language difficulties (e.g. ambiguity and vagueness), evaluating validity of individual arguments and truth of individual premises, formulating a conclusion from premises, and recognizing alternative inferences. Using this rubric, it takes an hour to score a single essay. Each of the following tests were used in one study only: The American Council of Education's Test of Critical Thinking, Inclination toward Methodological Criticism, Ability at Methodological Criticism, Creative Reasoning Test, Florida Taxonomy of Cognitive Behavior, and the Cornell Critical Thinking Test.

Just as there are a wide variety of instruments used to measure critical thinking, evaluations come from a wide range of disciplines and locations. (See column 2 of Table 2.) Four authors evaluate classes across disciplines. Others study effects of critical thinking programs on students from a single discipline including classes in mass communication, business, biology, and sociology.

What are relative merits for essay versus multiple choice tests for critical thinking? Some argue that essay tests are more valid because essay tests measure application of critical thinking skill, not merely knowledge of principles (Browne, Haas, Vogt, & West, 1977). While using the Watson-Glaser as their principal measure in a program at Bowling Green State University, Browne and his colleagues found that students, though able to demonstrate knowledge of critical thinking on the Watson-Glaser, still had trouble critically evaluating essays and other examples of thinking (Browne, Haas and Keeley, 1978). They argue that the multiple choice Watson-Glaser may measure the ability to recognize a valid syllogism, but may not test the ability of students to apply valid deductive reasoning to a problem (Browne, Haas & Keeley, 1978). Evenhandedly, Browne and his associates concede that multiple choice tests are easy to use, have national norms, and take less time to score than do essay tests (Keeley, Browne, & Kreutzer, 1982).

How reliable are tests of critical thinking? Reliability is vital to any evaluation, because consistent measures help to rule out sources of variation that can obscure real effects of educational programs. A rough rule of thumb for interpreting reliability coefficients is that the closer they approach one the better. Values equal to or exceeding .70 are generally acceptable.

Evaluators using multiple choice tests did not measure the reliability either of the Watson-Glaser or the Cornell by using data from subjects participating in their evaluations. However, the manual for the Watson-Glaser (1980) reports test-retest reliability ($r = .75$), alternate forms reliability ($r = .75$ for Form A and Form B), and split-half reliability coefficients

Table 1 Credibility of Studies

Author 1	Type of Subjects 2	Study Design 3	No. in Pretest 4	No. in Posttest 5
Baker, P.J. & Anderson, L.E., 1983	Students in three sections of a Social Problems course at Illinois State University	Three-group pretest-posttest	$N_1 = 20$ $N_2 = 22$ $N_3 = 14$	$N_1 = 20$ $N_2 = 22$ $N_3 = 13$
Browne, M.N., Haas, P.F., Vogt, K.E. & West, J.S., 1977	Treatment group was freshmen in special course. Comparison group was seniors in business major.	Two-group pretest-posttest	T = 21 Compar. = 40	T = 21 Compar. = 40
Givens, C.F., 1976	40 randomly selected faculty and their students in classes at 4 universities	One-group posttest only	None	40 class of students in 4 universities
Keeley, S.M., Browne, M.N., & Kreutzer, J.S., 1982	Students at a midwestern university	Posttest only with nonequivalent groups	500 freshmen 500 seniors	155 freshmen 145 seniors
Lehmann, I.J. & Dressel, P.L., 1963	Students at Michigan State University	One-group pretest-posttest	Freshmen 590 M 461 F	Freshmen 590 M 461 F Sophomore 235 M 189 F Junior 179 M 144 M
Logan, C.H. 1976	Students at 8 levels in a large university, and one course in critical thinking (all in sociology)	Two group pretest-posttest with non-equivalent groups, Five groups post-tested only	E group = 84; comparison groups = 102, 30	E group = 67, comparison groups = 144, 32, 36, 42, 18.
Meiss, G.T. & Bates, G.W., 1984	Students in an introductory class in mass communication	Three-group pretest-posttest	N = 102	$N_1 = 27$ $N_2 = 26$ $N_3 = 30$
Smith, D.G., 1977; and Smith, D.G., 1983 (for more detailed description of the study)	Students in 12 classes, where teaching critical thinking was not a specific goal, at a small liberal arts college	One group pretest-posttest (12 classes combined)	N = 210	N = 138
Statkiewicz, W.R. & Allen, R.D., 1983	One section of 112 General Biology students at West Virginia University	One-group repeated measures design (measures made at three times during the semester)	N = 48	N = 48

Random Selection 6	Random Assignment 7	Period of Follow-up 8	Credibility Index 9
R_S	R_A	One semester of class	34
R_S	R_A	Academic quarter	26
R_S (classes selected randomly)	R_A	No follow-up	50
R_S	R_A	none	41
R_S	R_A	1, 2, 3 years	30
R_S	R_A	One semester in experimental group.	26
R_S	R_A (treatment assigned randomly to classes)	15 weeks	46
R_S	R_A	One semester	16
R_S	R_A	One semester	42

Table 2 Measures of Critical Thinking and Findings

Author	Type of Subjects	Measures of Critical Thinking	Reliability of Critical Thinking Test	Validity of Critical Thinking Test
1	2	3	4	5
Baker, P.J. & Anderson, L.E., 1983	Students in 3 sections of a social problems course at Illinois State University	Creative Reasoning Test	Inter-rater r = .70, .93, .96, .74	
Browne, M.N., Haas, P.F., Vogt, K.E., & West, J.S., 1977	Treatment group was freshmen in special course. Comparison group was seniors in business major.	The principal measure was a rubic devised by the authors to grade essay tests, plus the Watson-Glaser and Cornell.	Graders scored 122 essay tests and agreed within one letter grade on all but 5 of the 122 tests.	Authors argue that the essay test measures applied critical thinking.
Givens, C.F. 1976	40 randomly selected faculty and their students in classes at 4 universities	Florida Taxonomy of Cognitive Behavior (FTCB)	85% agreement on items for independent raters	Items based on Bloom's Taxonomy of Education Objectives
Keeley, S.M., Browne, M.N., & Kreutzer, J.S., 1982	Students at a midwestern university	Rubric developed by the authors for grading essay tests	Interrater reliability = .90	Authors think multiple choice tests fail to measure ability to identify argument and to generate criticism
Lehmann, I.J. & Dressel, P.L., 1963	Students at Michigan State University	American Council of Education's Test of Critical Thinking		
Smith, D.G., 1977; and Smith, D.G., 1983 (for more detailed description of the study)	Students in 12 classes, where critical thinking was not a specific goal, at a small liberal arts college	Watson-Glaser Critical Thinking Appraisal		
Statkiewicz, W.R., & Allen, R.D., 1983	One section of 112 General Biology students at West Virginia University	Practice Exercises (a forced choice, "defend your choice," test developed by authors)		Exercise correlated (r = .56, .59, and .71) with course examination grade
Logan, C.H. 1976	Students at 8 levels in large university, and one course in critical thinking, (all in sociology)	Inclination Toward Methodological Criticism; Ability at Methodological Criticism.		Items exemplify common fallacies
Meiss, G.T. & Bates, G.W. 1984	Students in an introductory class in mass communication	Watson-Glaser Critical Thinking Appraisal		

Hyp. Tested 6	Stat. Tested 7	SE 8	Comments 9
Students will improve their critical thinking skills pre to posttest in three classes	Percent improved	63% had sig. to moderate gains. 82% has greater posttest scores	Tests were scrambled so raters did not know pretest from posttest when scoring
Those in a business and society cluster course will score higher on an essay test of CT ability than will seniors in a comparison group at posttest. Those in a business and society cluster will score significantly higher at posttest than they did at pretest.	F test for homogeneity of variance, t test for difference of means P< .005 P< .005	 SE = 1.48 SE = 1.49	The Watson-Glaser and Cornell test were dropped from the analysis because of difficulty in data collection and because 6 students scored in the 85th percentile at pretest on WG.
1) The average level of classroom discourse is on the lowest cognitive level (FTCB) 2) There is no difference for professor's nor student's FTCB score between course level (basic/advanced), subject area, time the class had been in session 3) Students in small classes had higher cognitive level (FTCB) than in larger classes	(See Comments) t compar. of average medians P<N.S. P<N.S. P<N.S. P< .003		Both professors and students had highest mean for item 5 of FTCB "Give A Specific Fact"
Seniors will score higher on forms P and C of an essay test of critical thinking than will freshmen	ANOVA P< .05	No standard deviation so the SE can't be computed	The authors think the statistically significant differences favoring seniors mask small absolute differences. They're concerned that 40-60% of seniors failed to provide a single example of a logical flaw, significant ambiguity, or misuse of data in a written passage.
There will be stat. significant pre-to-post differences at various levels in critical thinking among MSU students by academic year. During Freshmen year (m + F) p< .001 During Sophomore year (M + F) p< .001 During Junior year p = N.S. During Senior year p< .01(M), p< .001 (F)	t for paired samples	Freshmen M .83 F .65 Sophomore M .27 F .21 Junior M -.01 F .07 Senior M .12 F .15	The appendix does not contain a copy of the instrument. Sophomores and Juniors are selected randomly. Freshmen and Seniors are pre-post tested. SE appears to be most pronounced in first wo years.
The principal hypotheses concern interactions. The three teaching process variables that were associated significantly with Watson-Glaser class means were greater student participation in class discussion r = .63, P< .025; teacher encouragement r = .62, P< .025; higher peer to peer interaction r = .57, P< .05. A pre-post comparison of means for Watson-Glaser test was planned, but the comparison was not made because the means were almost identical.	Canonical correlation, bivariate correlation for interactions.	SE = 0 (approx.)	Classes with low participation tended to decline in critical thinking, suggesting to Smith that a decline in critical thinking may result in classes that emphasize memorizing and a lack of practice.
Consistent execution of practice exercises will lead to higher practice exercise scores in 12 member groups of randomly chosen A, B, C, D grade level students	ANOVA P< .009	Can't compute, (no standard deviation)	The author's inferences that the practice exercises are a highly productive component of the program are weakly supported due to a lack of control or even a comparison group.
Students who have more sociology courses will be more **inclined** to think critically. Students who have more sociology courses will be better able to think critically when they are **specifically instructed** to do so. Findings: Freshmen and Sophomores in a critical thinking course identified an average of 2.3 and 2.4 fallacies of 10 possible, more than in all eight levels of students including teaching assistants in other classes.	Pearson r = -.24 P< .01 Pearson r = .01, N.S.	Can't compute SE because no standard deviation given	Though students in the critical thinking group only spotted an average of 2.3 to 2.4 among 10 possible errors in thinking, the range for those in introductory ($\bar{x}$ = .29-.68) graduate students ($\bar{x}$ = 1.3-1.46) was lower than the mean in a special critical thinking course.
There will be statistically significant pre-post difference scores on WGCTA among Declarative Sentence Guide, Question Guide, and Topic outline (control group). Only group improved with stat. sig. was class with Declarative Sentence Guide	ANOVA t-test P = not reported	 SE = .63	A strength of this study is its random assignment of treatments to intact classes.

ranging from .69 to .89. Reliability for the Cornell (Ennis, Millman & Tomko, 1985) appears to be higher for Level X (the average for fourteen coefficients is .80) than for Level Y (the average for fourteen coefficients is .71). For those interested in a more detailed discussion of reliability for published tests, Ennis' (1984) critique will be helpful. Ennis says critical thinking may be multidimensional; so tests of reliability by measures of internal consistency may be misleading.

Those using essay tests did evaluate the reliability of instruments used in their evaluations. Browne and others (1977) found that, when graders scored 122 essay tests, scorers agreed within one letter grade on all but five essays. Their criteria for grading essays may have been honed to a finer edge, because their more recent study, using the same rubric for scoring, reported a .90 inter-rater correlation (Keeley, Browne, & Kreutzer, 1982). Baker and Anderson (1983) reported an average inter-rater correlation of .83 for their Creative Reasoning Test in a social problems course.

Givens (1976) reported 85% agreement between raters who applied the Florida Taxonomy of Cognitive Behavior (FATB) to audiotapes of classroom behavior. The FATB measures Bloom's taxonomy of educational objectives including the following low to high hierarchy: knowledge of specifics, translation, interpretation, application, analysis, synthesis, and evaluation.

Thus, it seems that multiple choice and essay tests for critical thinking can be scored reliably.

Next is evidence regarding the effectiveness of conventional curricula on critical thinking. Five studies evaluate standard curricula. Among these, four examine differences between advanced and less advanced students in the same university (Givens, 1976; Lehmann and Dressel, 1963; Logan, 1976; Keeley, Browne, & Kreutzer, 1982), and one study examines pre-post differences during one semester (Smith, 1977).

Givens (1976) randomly selected forty faculty from four universities to represent large and small, public and private institutions. Givens' survey revealed no statistically significant difference between basic and advanced university students on the Florida Taxonomy of Cognitive Behavior (FATB). Her most striking finding may be that student and faculty discourse on the FATB averaged on the lowest cognitive level (knowledge), but professors were slightly lower, on the average, than were their students. This may reflect faculty who lecture and students who ask questions about lecture content.

Lehmann and Dressel (1963) did a three-year longitudinal study of students at Michigan State University. They found a statistically significant improvement in critical thinking on freshman-to-sophomore, sophomore-to-junior, and junior-to-senior comparisons. Strength of effect is substantial for the freshman year but drops sharply thereafter. Several problems with the study make interpreting these findings difficult. Their large sample may inflate significance levels. The apparent improvement may reflect factors other than education, including maturation as students age, and effects of life experiences outside the university.

Logan (1976) tested students' Inclination toward Methodological Criticism (students were instructed to just react to a series of ten statements containing common fallacies in thinking about social issues); he also tested their Ability at Methodological Criticism (students were instructed specifically to think clearly and scientifically about each statement). He applied his measures to 874 sociology students at eight levels, from freshmen to graduate teaching assistants, at a large midwestern university. He found a negative correlation between number of sociology courses taken and inclination to think critically ($r=-.24, p<.01$)! He concluded, "One plausible explanation is that what a lot of sociologists say and what they do are often very

different things. The professed concern among sociologists with teaching students to think more rationally and scientifically about social phenomena may be to a considerable degree lip service that masks a hidden curriculum. Sociology professors may in fact be more concerned with teaching students what to think than how to think."

Keeley and others (1982) randomly selected 500 seniors and 500 freshmen (they got responses from 155 freshmen and from 145 seniors) among students at a midwestern university. They administered a reliable ($r=.90$) essay test to both groups. Seniors did statistically significantly better on the test, but Keeley and his associates considered performance to be disappointingly low for both groups. Across the items on the test, an average of 51% of freshmen and 42% of seniors got no points for items on the essay.

Smith (1983) reported no difference over one semester for the Watson-Glaser. He pretested and posttested students in 12 classes at a small liberal arts college.

The preceding results seem to show that conventional curricula, not designed specifically to teach critical thinking, may produce weak positive effects, no effect, or even harmful effects on critical thinking. However, these findings are hard to interpret. An apparent improvement maybe due to normal maturation of students during their college careers, possibly because student drop-out or "mortality" may leave more competent students to take later measures, thus giving an illusion of an educational effect. Events other than those taking place in the university experience may change thinking ability. Without random assignment to educational programs, such alternate explanations for findings are numerous (Campbell & Stanley, 1963; Cook and Campbell, 1979).

Which faculty and student behaviors are most associated with learning critical thinking? Such associations are especially important because they may suggest ways to design successful programs. Below are findings from three studies giving information about factors associated with students' learning critical thinking.

Smith (1977) reports statistically significant associations between high Watson-Glaser scores and greater student participation in class discussion ($r=.63$, $p<.025$), higher encouragement by the teacher ($r=.62$, $p<.025$), and higher peer-to-peer interaction ($r=.57$, $p.<.05$). Givens (1976) found that scores on the Florida Taxonomy of Education Objectives were higher for students in small classes and higher in large institutions. She also found a positive association between "analysis" by professors and performance at that level by students ($r=.18$, $p<.03$, $N=155$), but no consistent relationship between cognitive level of professors and corresponding cognitive level of students in their classes. Givens also found no significant difference in cognitive level of discourse between professors and students by type of institution (public or private), course level (beginning or advanced), subject area, time the class had been in session, or within or between institutions. Statkiewicz and Allen (1983) found that biology students who did exercises designed to force them to make choices did better on the final course grade.

Though studies of association suggest factors that might be harnessed to drive a critical thinking program toward its goals, association evidence is weak: characteristics of the learning environment that seem to affect critical thinking may themselves be only associated with real causal factors. For example, peer-to-peer interaction may be associated with better performance, but itself may be only a reflection of some particular feature of a well-conducted discussion.

How effective are special courses designed specifically to teach critical thinking? Four studies address this question. All apply critical thinking to issues in the social sciences.

Baker and Anderson (1983) think

most social problems courses merely teach students to memorize and recall: such courses do not teach students to critically examine social issues. Baker and Anderson teach their students to scan the popular press to identify a problem commonly discussed there; define the problem; stipulate its various causes, and offer general and specific solutions. They developed a Creative Reasoning Test to measure their intended goals and used it to evaluate the effects of three different teaching methods. Their Structured Inquiry method (where specific learning goals are set for each student around some analytical thinking skill) produced the highest percentage of gain, but Focused Inquiry (where students select a topic, gather literature about it, and design a study), and Open Ended Inquiry (where students compared two different modes of inquiry including journalistic and sociological approaches) also produced substantial percentage gains.

Meiss and Bates (1984) evaluated three methods for teaching critical thinking in an introductory mass communications class. Their methods included: a manual by Meiss employing declarative sentences to help students to use synthesis, application, and evaluation; the same manual used to pose thought-provoking questions, and a control group who merely got a topic outline. These methods were randomly assigned to three classes who attended the same lecture but were exposed to different methods in each quiz section. Analysis of variance comparisons were done at the end of the fifteenth week of the semester. The only statistically significant improvements on the Watson-Glaser were among students exposed to the declarative sentence method.

I was a student in the experimental course that Logan (1976) evaluated. The instructor for the course, Professor Michael Hakeem, used no text. He read aloud parts of the day's readings and "thought out loud" about the readings for the benefit of the class. Students in his class were given the chance to read critically and react aloud to thinking in professional articles, books, and stories in the popular press. He criticized these student reactions for the benefit of the class. We were encouraged to think about the method by which each author drew conclusions, and not get too involved in the content of the material. Tests were short essay based on readings and ideas given to us. Most failed the first essays because we merely parroted back the test material—an effective procedure in most other classes.

According to Logan (1976), those who took this experimental freshman and sophomore course were able to spot an average of 1.79 fallacies among a possible ten on a scale measuring inclination to think scientifically; they spotted 2.35 when told specifically to think scientifically. Not bad compared with graduate teaching assistants in the same department who scored 1.11 and 1.92 respectively.

The fourth and final evaluation of a course specifically designed to teach critical thinking is one by Browne and others (1977). They developed a new freshman-level business course. Its objectives were: developing critical thinking skills, developing respect for alternate viewpoints, and generating alternate hypotheses. They scored the essays of freshmen in the Business and Society Cluster course and concurrently scored essays done by a comparison group of senior business majors. Freshmen out performed seniors at posttest. Pretest scores were almost identical for freshmen and seniors. Of this they say, "This [no difference] result was surprising because we had expected the seniors to perform significantly better than the cluster [freshmen] students at pretest. Some further examination appears to be necessary to determine whether the application of critical skills is actually assimilated during a traditional four-year curriculum."

Conclusions

The four studies reported immediately above seem to indicate that critical thinking can be effectively taught at the university level. However, a caution is warranted. Not a single study among the nine reported here used random assignment to treatment groups nor to treatment and control groups. Thus, inferences about the effects of university teaching on critical thinking must be made with caution.

It is not surprising that critical thinkers would omit a major criterion for making casual inferences when designing their experiments. It is like pulling teeth to extract data from busy faculty, especially controls. Our recent experience with a randomized study of critical thinking at the University of Wisconsin—Eau Claire has made us much more appreciative of the studies reported here and respectful of problems with randomized trials of new programs.

Students who have particular abilities, cognitive styles, experiences, and levels of motivation may benefit best from particular teaching approaches. But no aptitude-by-treatment interaction (ATI) studies were found. Those who want to evaluate critical thinking programs from ATI perspective might base their procedures and hypotheses on ATI research in science teaching (Koran & Koran, 1984) and discussions of how to design such research (Cronbach & Snow, 1977).

The critical thinking movement seems to be gathering momentum. Recently, journals have devoted whole issues to teaching critical thinking (See the **National Forum** for Winter 1985) educators have initiated compulsory tests for critical thinking statewide (Kneedler, 1985), and approximately nine hundred attended the Third International Conference on Critical Thinking and Educational Reform, where the conference director had conservatively expected from four to five hundred (Paul, 1985; R.W. Paul in a personal communication, June 23, 1986).

Educators involved in the critical thinking movement might be able to direct their efforts more effectively if more research were available to guide them. Such research might be more useful if randomized studies were available to evaluate aptitude-by-treatment interactions, the relative merits of different teaching approaches, various aspects of the classroom environment including, for example, class size, and researchers met to isolate major dimensions of critical thinking and standardized measures for those dimensions.

References

Baker, P. J., & Anderson, L. E. **Teaching social problems through critical reasoning.** Washington D.C.: Teaching Resource Centre, 1983. (ERIC Document Reproduction Service No. 013371)

Baker, P. J. "Learning sociology and assessing critical thinking." In P. J. Baker, & E. K. Wilson (Eds.), **Knowledge available and knowledge needed to improve instruction in sociology**. Washington D.C.: American Sociological Association, 1979.

Browne, M. N., Haas, P. F., & Keeley, S. "Measuring critical thinking skills in college." **Educational Forum**, 1978, 42(2), 219-226.

Browne, M. N., Haas, P. F., Vogt, K. E., & West, J. S. "Design and implementation of an evaluation procedure for an innovative undergraduate program." **College Student Journal**, 1977, 11 (4), Pt. 2, 1-10.

Campbell, D. T. & Stanley, J. C. **Experimental and quasi-experimental designs for research**. Boston: Rand McNally, 1963.

Cook, T. D., & Campbell, D. T. **Quasi-experimentation: Design & analysis issues for field settings**. Boston: Houghton Mifflin, 1979.

Cronbach, L. J., & Snow, R. E. **Aptitudes and instructional methods.** New York: Irvington Pub., 1977.

Ennis, R. H. "Problems in testing informal logic: critical thinking,

reasoning ability." **Informal Logic**, 1984, 6(1), 3-9.

Ennis, R. H., Millman, J. & Tomko, T. N. **The Cornell critical thinking tests level X & level Z-Manual.** Midwest Publications: Pacific Grove, CA, 1985.

Gibbs, L.E. **Quality of Study Rating Form**. Department of Social Work, University of Wisconsin—Eau Claire, Eau Claire, WI, 1985.

Givens, C. F. **A descriptive study of the cognitive level of classroom discourse of college professors and students.** Unpublished doctoral dissertation, Clairemont Graduate School, 1976.

Glass, G. V. "Meta-analysis: An approach to the synthesis of research results." **Journal of Research in Science Teaching**, 1982, 19(2), 93-112.

Keeley, S. M., Browne, M. N., & Kreutzer, J. S. "A comparison of freshmen and seniors on general and specific essay tests of critical thinking." **Research and Higher Education**, 1982, 17(2), 139-154.

Hedges, L. V. "Advances in statistical methods of meta-analysis." In I. W. H. Yeaton, & P. M. Wortman (Eds.), **Issues in data synthesis. New directions for program evaluation series No.24**, San Francisco: Jossey-Bass, 1984.

Kneedler, P. **California's assessment of the critical thinking skills in history-social science.** Unpublished manuscript, 1985. (Available from California Assessment Program, 721 Capitol Mall, Sacramento, California 95814.)

Koran, M. L., & Koran, J. J. "Aptitude-treatment interaction research in science education." **Journal of Research in Science Teaching,** 1984, 21 (8), 793-808.

Lehmann, I. J., & Dressel, P. L. **Changes in critical thinking ability, attitudes, and values associated with college attendance**. Final report of cooperative research project #1646, East Lansing, Michigan: Michigan State University, 1963.

Logan, C. H. "Do sociologists teach students to think more critically?" **Teaching Sociology,** 1976, 4 (1), 29-48.

Meiss, G. T. & Bates, G. W. **Cognitive and attitudinal effects of reasoning message strategies**. Paper presented at the thirty-fourth annual meeting of the International Communications Association, San Francisco, California, May 24-28, 1984. (ERIC Document Reproduction Series No 246514)

Noyce, R.M. **An experiment in developing critical thinking abilities through persuasive communication**. Unpublished doctoral dissertation, University of Missouri-Kansas City, 1970.

Norris, S. P. "Synthesis of research in critical thinking." **Educational Leadership**, 1985, 42 (8), 40-45.

Paul, R. W. "Critical thinking: The state of the field." Paper presented at the Third International Conference on Critical Thinking and Educational Reform, Sonoma State University, Rhonert Park, California, July 1985.

Smith, D. G. "College classroom interactions and critical thinking." **Journal of Educational Psychology**, 1977, 60 (2), 180-190.

Smith, D. G. "Instructions and outcomes in an undergraduate setting." In C. L. Ellner & C. P. Barnes (Eds.), **Studies in college teaching: Experimental Results**. Lexington, Mass.: DC Heath and Company, 1983.

Smith, E.R. & Tyler, R.W. **Appraising and recording student progress.** New York: Harper & Brothers, 1942.

Statkiewicz, W. R., & Allen, R. D. "Practice exercises to develop critical thinking skills." **Journal of College Science Teaching**, 1983, 12 (4), 262-265.

Watson, G. & Glaser, E. M. **Watson-Glaser critical thinking appraisal.** New York: The Psychological Corporation, 1980.

Note

*If you are conducting an evaluation or know of one, please let me know. I am happy to send reprints of this article. The author acknowledges comments and suggestions by Michael Hakeem, Pat Kark, John Morris, Diana Sigler and Michael Stratton. This review was supported by funds from the Undergraduate Teaching Council, the University of Wisconsin System, and by the Office of Graduate Studies and University Research, University of Wisconsin—Eau Claire.

Dr. Leonard E. Gibbs, Department of Social Work, University of Wisconsin—Eau Claire, Eau Claire, WI 54701 □