The effectiveness of computer-assisted instruction in critical thinking

David Hitchcock

Department of Philosophy, McMaster University

Hamilton, Canada L8S 4K1

hitchckd@mcmaster.ca

The effectiveness of computer-assisted instruction in critical thinking

1. Introduction

Undergraduate critical thinking courses are supposed to improve skills in critical thinking and to foster the dispositions (i.e. behavioural tendencies) of an ideal critical thinker. Students taking such courses already have these skills and dispositions to some extent, and their manifestation does not require specialized technical knowledge. Hence it is not obvious that a critical thinking course actually does what it is supposed to do. In this respect, critical thinking courses differ from courses with a specialized subject-matter not previously known to the students, e.g. organic chemistry or ancient Greek philosophy or eastern European politics. In those courses performance on a final examination can be taken as a good measure of how much a student has learned in the course. In a critical thinking course, on the other hand, a good final exam will not be a test of such specialized subject-matter as the construction of a Venn diagram for a categorical syllogism or the difference between a reportive and a stipulative definition, but will ask students to analyze and evaluate, in a way that the uninitiated will understand, arguments and other presentations of the sort they will encounter in everyday life and in academic or professional contexts. Performance on such a final examination may thus reflect the student's skills at the start of the course rather than anything learned in the course. If there is improvement, it may be due generally to a semester of engagement in undergraduate courses rather than specifically to instruction in the critical thinking course. There may even be a deterioration in performance from what the student would have shown at the beginning of the course.

We therefore need well-designed studies of the effectiveness of undergraduate instruction

in critical thinking, whether in stand-alone courses or infused into disciplinary courses (or both). There is a particular need to compare the effectiveness of different forms of instruction in critical thinking. With the widespread diffusion of the personal computer, and financial pressures on institutions of higher education, instructors are relying more and more on drill-and-practice software, some of which has built-in tutorial helps. This software can reduce the labour required to instruct the students; at the same time, it provides immediate feedback and necessary correction in the context of quality practice, which some writers (e.g. van Gelder 2000, 2001) identify as the key to getting substantial improvement in critical thinking skills. Does the use of such software result in greater skill development, less, or about the same? Can such software completely replace the traditional labour-intensive format of working through examples in small groups and getting feedback from an expert group discussion leader? Or is it better to combine the two approaches? Can machine-scored multiple-choice testing completely or partially replace human grading of written answers to open-ended questions? Answers to such questions can help instructors and academic administrators make wise decisions about formats and resources for undergraduate critical thinking instruction. We do not have those answers now.

An ideal design for a study of a certain method of teaching critical thinking would take a representative sample of the undergraduate population of interest, divide it randomly into two groups, and treat the two groups the same way except that one receives the critical thinking instruction and the other does not. Each group would be tested before and after the instructional period by some validated test of the outcomes of interest. If statistical analysis shows that the mean gain in test scores is significantly greater in the group receiving critical thinking instruction than in the control group, then the critical thinking instruction has in all probability achieved the desired

effect, to roughly the degree indicated by the difference between the two groups in mean gains. Alternatively, a representative sample of undergraduate students could be randomly allocated to one of two (or more) groups receiving instruction in critical thinking, with the groups differing in the method of instruction, learning and testing. Statistically significant differences between the groups' mean gains would indicate that one method was more effective than another. For either type of study, statistically significant differences are not necessarily educationally meaningful; with large groups, even slight differences will be statistically significant, but they will not reflect much difference in educational outcome. Judgement is required to determine how much of a difference is educationally meaningful or important. A useful rule of thumb is that a medium effect size is a difference of 0.5 of a standard deviation in the population (Cohen 1998: 24-27); Norman et al. (2003) report that minimally detectable differences in health studies using a variety of measurement instruments average half a standard deviation (mean = 0.495SD, standard deviation = 0.155), a figure which can be explained by the fact, established in psychological research, that over a wide range of tasks the limit of people's ability to discriminate is about 1 part in 7, which is very close to half a SD.

Practical constraints make such ideal designs impossible. Students register in the courses they choose, and cannot reasonably be forced by random allocation either to take a critical thinking course or to take some placebo. The only practically obtainable control group is a group of students who have a roughly similar educational experience except for the absence of critical thinking instruction; practically speaking, one cannot put together a group taking exactly the same courses other than the critical thinking course. Further, there are disputes about the validity of even standardized tests of critical thinking skills. And, although there is one standardized test of critical thinking dispositions (the California Critical Thinking Disposition Inventory), questions can be raised about how

accurately students would answer questions asking them to report their attitudes; self-deception, lack of awareness of one's actual tendencies and a desire to make oneself look good can all produce inaccurate answers.

A standard design therefore administers to a group of students receiving critical thinking instruction a pre-test and a post-test using a validated instrument for testing critical thinking skills. Examples of such designs are studies by Facione (1990a), Hatcher (1999), van Gelder (2000, 2001) and Twardy (forthcoming). All four of these studies used the California Critical Thinking Skills Test (CCTST) developed by Facione (Facione et al., 1998), thus facilitating comparison. Facione's study included a control group of 90 students in an Introduction to Philosophy course, whose mean gain in CCTST score can thus be used as a basis of comparison. Since no study of the effect of critical thinking instruction has used a randomized experimental design, with subjects randomly allocated to an intervention group and a control group otherwise treated equally, there is no true control group. The gains reported for different course designs offer a relative comparison, rather than an absolute measure of effect size.

The present study used a similar design to determine the gain in critical thinking skills among a group of undergraduate students whose instruction in critical thinking completely replaced face-to-face tutorials with computer-assisted instruction with built-in tutorial helps, and whose grade depended entirely on multiple-choice testing. Such a course design is remarkably efficient, but how effective is it? That is what this study tried to determine, at least for critical thinking skills.

2. Method

402 undergraduate students at McMaster University in Hamilton, Canada completed a 13-week course in critical thinking between January and early April of 2001, meeting in one group for two 50-minute classes a week. At the first meeting the course outline was reviewed and a pre-test announced, to be administered in the second class; students were told not to do any preparation for this test. Those students who attended the second class wrote as a pre-test either Form A or Form B of the California Critical Thinking Skills Test (CCTST). The following 11 weeks were devoted to lectures about critical thinking, except that two classes were used for in-class term tests and one class was cancelled. Thus the students had the opportunity to attend 19 lectures of 50 minutes each, i.e. to receive a total of 15.8 hours of critical thinking instruction. In the 13th week, those students who attended the second-last class of the course wrote as a post-test either Form A or Form B of the CCTST. The last class was devoted to a review of the course and an explanation of the format of the final examination.

There were no tutorials. Two graduate teaching assistants and the instructor were available for consultation by e-mail (monitored daily) or during office hours, but these opportunities were used very little, except just before term tests; the course could have been (and subsequently was) run just as effectively with one assistant. Review sessions before the mid-term and final examination were attended by about 10% of the students. Two assignments, the mid-term test and the final examination were all in machine-scored multiple-choice format; in other words, there was no written graded work.

Students used as their textbook Jill LeBlanc's *Critical Thinking* (Norton, 1998), along with its accompanying software called LEMUR, an acronym for Logical Evaluation Makes Understanding Real. The course covered nine of the 10 chapters in the book and accompanying software, with the

following topics: identifying arguments, standardizing arguments, necessary and sufficient conditions, language (definitions and fallacies of language), accepting premises, relevance, arguments from analogy, arguments from experience, causal arguments. There were two multiple-choice assignments, one on distinguishing arguments from causal explanations and standardizing arguments,[1] the other on arguments from analogy. The mid-term covered the listed topics up to and including accepting premises. The final exam covered all the listed topics. The software LEMUR has multiple-choice exercises and quizzes tied to the book's chapters, with tutorial help in the form of explanations and hints if the user chooses an incorrect answer; if the user answers an item correctly, there is often an explanation why that answer is correct. The software includes pre-structured diagrams into which students can drag component sentences of an argumentative text to note its structure, but does not allow the construction of original diagrams; in this respect it is less sophisticated than Athenasoft (available at www.athenasoft.org), Araucaria (available at http://www.computing.dundee.ac.uk/staff/creed/araucaria/download.html), and Reason!Able (available at http://www.goreason.com/download.htm).There was a Web site for the course, on which answers to the textbook exercises were posted, as well as past multiple-choice assignments, tests and exams with answers, along with other helps. There was no monitoring of the extent to which a given student used the software or the Web site.

To encourage students to do their best on both the pre-test and the post-test, 5% of the final grade was given for the better of the two marks received; if one of the two tests was not written the score on the other test was used, and if neither test was written the final exam counted for an additional 5%. In accordance with the test manual, students were not told anything in advance about the test, except that it was a multiple-choice test. A few students who asked what

they should do to study for the post-test were told simply to review the material for the entire course. Students had about 55 minutes on each administration to answer the items, slightly more than the 45 minutes recommended in the manual.

The original intention was to use a simple crossover design, with half the students writing Form A as the pre-test and Form B as the post-test, and the other half writing Form B as the pre-test and Form A as the post-test. This design automatically corrects for any differences in difficulty between the two forms. As it turned out, far more students wrote Form A as the pre-test than wrote Form B, and there were not enough copies of Form B to administer it as a post-test to those who wrote Form A as the pre-test. Hence the Form A pre-test group was divided into two for the post-test, with roughly half of them writing Form B and the rest writing Form A again. This design made it possible to determine whether it makes any difference to administer the same form of the test as pre-test and post-test, as opposed to administering a different form.

3. Results

3.1 Mean gain overall: Of the 402 students who completed the course, 278 wrote both the pre-test and the post-test. Their mean score on the pre-test was 17.03[2] out of 34, with a standard deviation of 4.45. Their mean score on the post-test was 19.22 out of 34, with a standard deviation of 4.92. Thus the average gain was 2.19 points out of 34, or 6.44 percentage points (from 50.08% to 56.52%). The mean difference in standard deviations, estimating the standard deviation in the population at 4.45, is .49.[3] The difference is statistically significant (p=.00),[4] and is substantially greater than the difference of .63 points out of 34, or .14 standard deviations,

reported for a control group of 90 students taking an introductory philosophy course (Facione 1990a: 18). Results for the 278 McMaster students, for the control group, and for groups taking critical thinking courses elsewhere are recorded in Table 1.

[insert Table 1 about here]

3.2 Mean gain by form type: The 278 students fell into four groups, according to which form of the test they wrote on the pre-test and post-test. I designate these groups "AB", "AA","BA"and "BB", with the first letter indicating the form written as a pre-test and the second the form written as a post-test. The mean score of the 90 students in group AB increased from 17.34 out of 34, with a standard deviation of 4.59, to 19.22, with a standard deviation of 4.75; the AB group's average gain was thus 1.88 points out of 34, or .42 of the estimated standard deviation in the population. The mean score of the 79 students in group AA increased from 16.45 out of 34, with a standard deviation of 4.30, to 18.56, with a standard deviation of 4.94; the AA group's average gain was thus 2.11 points out of 34, or .47 of the estimated standard deviation in the population. The mean score of the 108 students in group BA increased from 17.20 out of 34, with a standard deviation of 4.45, to 19.73, with a standard deviation of 5.04; the BA group's average gain was thus 2.53, or .56 of the estimated standard deviation in the population. There was only one student in group BB; his score was 17 out of 34 on both the pre-test and the post-test. The results are consistent with form B being slightly more difficult than form A, since there was more improvement in going from form B to form A than vice versa, and an intermediate degree of improvement in those writing form A twice.[5] But the differences in improvement by form type

are not statistically significant (p=.45).[6] The intermediate gain by the group which wrote form A twice indicates no trace of a difference between writing the same form of the test twice, as opposed to writing a different form in the post-test; this result confirms that reported in the test manual: "We have repeatedly found no test effect when using a single version of the CCTST for both pre-testing and post-testing. This is to say that a group will not do better on the test simply because they have taken it before." (Facione et al. 1998: 14) Table 2 shows the results for the 278 students as a whole and for each sub-group by form type. Figure 1 displays the mean gain, expressed as a percentage of the estimated standard deviation in the population, for the whole group and for each of the three sub-groups by form type.

As explained in note 5, allocation of students at pre-test was not random. As it turned out, the group writing form B at pre-test did better on average than the group writing Form A. Even though Form B is more difficult than Form A (as indicated in the previous paragraph, cf. Jacobs 1995, 1999), the mean score at pre-test on Form B (17.38 ± 4.54) was slightly higher than that on Form A (16.96 ± 4.47) , as indicated in Table 2. The difference was not statistically significant.[7]

[insert Table 2 about here]

[insert Figure 1 about here]

Not all students who wrote the pre-test also wrote the post-test. Some dropped the course, while others simply chose not to use precious time in the last week of term trying to improve (marginally, as it turned out) a score which was worth only 5% of the final grade. It is therefore

worth asking whether the students who took both tests differed in their mean score from the students who took only the pre-test. The short answer is: no. As reported above, the pre-test mean score among the 278 students who also took the post-test was 17.03, with a standard deviation of 4.45. The pre-test mean score among the 96 students who did not take the post-test was 17.35, with a standard deviation of 4.63. The difference was not statistically significant (p=.55).[8] Thus there is no reason to believe that the 96 students would have shown significantly different mean gains if they had written the post-test after having taken the course.

3.3 Mean gain by Faculty of registration: Students registered in the course came from programs in five different Faculties of the university: business, engineering, humanities, science, social sciences.[9] Among the 278 students who wrote both the pre-test and the post-test, there were slight but statistically significant differences (p=.02) in the mean pre-test score by Faculty, which ranged from a low of 16.07 for the engineering students to a high of 18.34 for the science students.[10] But each of the five groups showed an improvement in the post-test, and the differences in mean gain were not statistically significant (p=.34), partly because there were very few students from some Faculties.[11] For details, see Table 3.

[insert Table 3 about here]

3.4 Mean gain by level of registration: Students registered in the course were predominantly sophomores (Level 2 students), who would typically be 20 years old, given that the educational jurisdiction from which the vast majority of these students came (the Canadian province of Ontario) at the time required five years of secondary school education for university entrance and

children in this jurisdiction normally start secondary school in the calendar year in which they turn 14. But there were also a substantial number of juniors (Level 3 students), as well as smaller numbers of seniors (Level 4 students) and students registered in Level 1, as well as three students registered in Level 5 of a 5-year program. Thus it was possible to see whether initial performance and mean gains differed by the stage of a student's university education. Among the students who wrote both the pre-test and the post-test, there was a slight but statistically significant difference in initial performance by Level of registration, mainly due to a substantially lower mean pre-test score among the 23 Level 1 students (14.52) than among those in Levels 2 and above, whose mean pre-test score ranged from 17.05 among the 158 Level 2 students to 17.65 among the 79 Level 3 students.[12] Almost all the students registered in Level 1, however, were in their second year of full-time undergraduate work; they had apparently failed to complete a full complement of courses for their Level I program in their first year of university, and would thus be a weaker group of students.[13] The mean pre-test score of the Level 3 students was only .13 standard deviations (.60 points out of 34) above that of the Level 2 students, and the mean pre-test score of the Level 4 and 5 students was only .08 standard deviations (.36 points out of 34) above that of the Level 2 students. Although these differences would be affected by differences among the groups with respect to verbal and mathematical aptitude, and similar causally relevant factors, the small size of the differences is consistent with findings in other studies that, in the absence of instruction dedicated to developing critical thinking skills, merely taking university courses produces little improvement in critical thinking skills after the freshman year. Students at each of the four levels, however, showed gains on the post-test, which suggests that a stand-alone course in critical thinking produces improvements in critical thinking skills which the students would

not have acquired simply by taking another year of university courses. The differences in mean gain by Level were not statistically significant (p=.80).[14] When the mean gain by Level was expressed as a percentage of the standard deviation on the pre-test, the range was quite narrow, from a low of 0.43 SD for the Level 1 students to a high of 0.62 SD for the Level 4 and 5 students. For details, see Table 4.

[insert Table 4 about here]

3.5 Mean gain by type of item: Not all the skills specifically tested in the CCTST were specifically taught in the course. In fact, the course followed a textbook and software which had been designed independently of the CCTST and of the expert consensus on critical thinking on which it is based (American Philosophical Association 1990). An example of non-overlap is a set of items on the CCTST which require evaluation of inferences in categorical syllogisms; the course taken by students in the present study skipped the textbook's chapter on categorical syllogisms. In order to determine which items on the CCTST tested items specifically taught in the course, five people (the instructor [i.e. the author of the present report], two teaching assistants in the course, the instructor in the evening section of the same course, and a graduate student who had previously assisted in the course) independently classified each of the 34 items as testing a skill (a) definitely taught in the course, (b) definitely not taught in the course, or (c) neither definitely taught nor definitely not taught (i.e. borderline). Inter-rater agreement on the classification was surprisingly low, with a reliability coefficient of .50 and .45 for Forms A and B respectively; disagreements on how to classify items can perhaps be explained by differences in

the level of abstraction at which raters classified the skill tested by each item (e.g. as determining whether a given sentence follows necessarily from given sentences or as determining whether a given categorical syllogism is deductively valid). Since the items on the two forms exactly parallel one another, overall classification of an item by the five raters was determined by giving each numbered item (on either form) 3 points for each rating as definitely taught, 2 points for each borderline rating, and 1 point for each rating as definitely not taught. This produced for each of the 34 items a total score ranging from 10 (uniformly rated as definitely not taught in the course) to 30 (uniformly rated as definitely taught in the course). Items with total scores over 25 were classified as definitely taught in the course; there were 25 such items out of the 34 (numbers 1-4, 6, 8, 10-15, 20-22, 24-28 and 30-34). Items with total scores between 15 and 25 were classified as possibly taught in the course; there were 7 such items out of the 34 (numbers 5, 7, 9, 16, 19, 23 and 29). Items with total scores below 15 were classified as definitely not taught; there were 2 such items out of the 34 (17 and 18). Rather surprisingly, the mean gain on each class of items was almost exactly the same when expressed as a percentage of the total possible score: 6.44% of the total possible score (1.61 out of 25) for items testing skills definitely taught in the course, 6.71% of the total possible score (0.47 out of 7) for items testing skills possibly taught in the course, and 6.00% of the total possible score (0.12 out of 2) for items testing skills definitely not taught in the course. With each of the three groups of items, the gain in mean score was statistically significant ($p=.00$).[15] The close similarity in the percentage gain among different types of items suggests that, if the improvement from pre-test to post-test is due to the course (as the comparison with the control group would indicate), it is due to rather general features of the students' learning rather than to such details as whether they specifically learned categorical

syllogisms.

[insert Table 5 about here]

4. Discussion

In one semester, the 278 students in this study improved their score on the California Critical Thinking Skills Test (Forms A and B) by an average of 2.19 points out of 34, a gain of 6.44 percentage points from 17.03 (50.08%) to 19.22 (56.52%). The mean gain was substantially greater than the gain by a control group of 90 students who took an introductory philosophy course at a California state university in early 1990; in one semester, that group of 90 students improved their score on the CCTST by an average of only 0.63 points out of 34, a gain of only 1.85 percentage points from 15.72 (46.23%) to 16.35 (48.08%). The difference of 1.56 points out of 34 (4.59 percentage points) is undoubtedly statistically significant (though no direct calculation of its level of significance was made), and is educationally meaningful. But it is not very impressive, especially when measured against the room for improvement which the McMaster students manifested on the pre-test. It is considerably less than gains ranging from 3.25 to 3.98 points out of 34 (9.55 to 11.70 percentage points) reported for three groups of students at the University of Melbourne after a one-semester course in critical thinking which combined computer-assisted instruction with graded written assignments and tests. On the other hand, the gain by the McMaster students is somewhat more than the mean gain of 1.44 points out of 34 (4.23 percentage points) reported for a group of 262 students at a California state university after

a one-semester course in critical thinking which did not use computer-assisted instruction.[16] For details and further comparisons with other studies, see Table 1.

To facilitate comparison with results of studies using other tests with different scoring systems, a standard measure of effect size is Cohen's d, i.e. the mean gain divided by the standard deviation. From this figure must be subtracted the mean gain in pre-test standard deviations which would have been expected without the critical thinking course. To arrive at this expected mean gain, we need to look at a synthesis of recent studies.

Pascarella and Terenzini (forthcoming) estimated on the basis of a synthesis of studies of U.S. students during the 1990s that college seniors had on average a critical thinking skills advantage over incoming freshmen of about .50 of a standard deviation. This estimate was substantially lower than their estimate in (Pascarella and Terenzini 1991) of a 1 standard deviation difference. In their new work, however, they estimate (Pascarella and Terenzini, forthcoming) that the first three years in college provide an improvement in critical thinking skills of about .55 of a standard deviation, an estimate which suggests that the estimate of .50 is too low. The 1990s studies which they reviewed found that most of the gains in critical thinking skills occur in a student's first year of college; they estimated that the sophomore advantage over freshmen was .34 of a standard deviation, the junior advantage over freshmen was .45 of a standard deviation, and the senior advantage over freshmen was .54 of a standard deviation. If we project a further .1 of a standard deviation as a result of the senior year in college, we arrive at an expected advantage of graduating seniors over incoming freshmen of .64 of a standard deviation, even if the undergraduate program does not include a stand-alone course in critical thinking. Assuming that in each of the four years of college, the gain is divided evenly between two

semesters, the expected gain over one semester for a freshman (Level 1 student) would be .17 of a standard deviation, and for a non-freshman (student in Level 2 or above) .05 of a standard deviation. The expected mean gain of .17 SD for freshmen compares quite well with the mean gain of .14 SD in CCTST scores reported by Facione (1990a: 16) in 90 students registered in a first-year Introduction to Philosophy course at a California state university, who were presumably mostly freshmen. Facione notes, however, that these 90 students, who got no credit in the course for their scores on either the pre-test or the post-test, seemed to take the task more seriously at the beginning of the term when they were fresh than another group of students had at the end of the previous term when they were exhausted; also, the 90 students were not told in advance that one of their last classes in the course would be devoted to writing a test which did not count towards their grade, a fact which may have led some of them to put in only a token performance. Hence Facione's results may understate the improvement by these 90 students in the skills tested by the CCTST. The present study found anomalously low pre-test means among the seniors (Level 4 and 5 students); although the juniors were .13 of a standard deviation above the sophomores, about as expected, the seniors were .05 of a standard deviation behind the juniors and only .08 of a standard deviation above the sophomores (Table 4). As mentioned above, the small number of Level 1 students in the present study (23) were mostly not freshmen, but sophomores who had not managed to complete their Level 1 program in their first year of study; hence their scores cannot be taken as any indication of scores on the CCTST that one would expect from McMaster freshmen. Further, the expected gain in critical thinking skills for such students after one semester of full-time undergraduate study without a critical thinking course should be about .05 of a standard deviation, since they are mostly not freshmen.

Subtracting this very rough estimate of .05 of a standard deviation from the gains shown by the students in the present study indicates that the course raised the critical thinking skills of the Level 1 students by about .38 of a standard deviation, of the Level 2 students by about .44 of a standard deviation, of the Level 3 students by about .41 of a standard deviation, and of the Level 4 students by about .62 of a standard deviation. In terms of effect sizes reported generally in education and the social sciences, the effect at each of the four levels of registration is between small and medium-sized (Cohen, 1988: 24-27). In terms of what one can reasonably expect from a one-semester course in critical thinking instruction, it seems intermediate; students in the one-semester stand-alone critical thinking courses tested by Facione (1990a) showed a gain of only .33 of a standard deviation, indicating that the critical thinking instruction caused a gain of about .16 of a standard deviation among the freshmen and about .28 of a standard deviation among the non-freshmen,[17] a small effect. On the other hand, Donohue et al. (2002) report mean gains ranging from .73 to .89 of a standard deviation among groups of freshmen in a one-semester course, for a net effect size ranging from .56 to .72 of a standard deviation, which is medium-sized to moderately large. In the present study the mean gain for the sophomores, juniors and seniors was in each case greater than they would be expected to show for the rest of their academic career without a course in critical thinking instruction; since this course was offered in the second semester, the sophomores would be expected without a critical thinking course to gain .25 of a standard deviation in the rest of their university career (vs. .49 actually gained after this one-semester course), the juniors .15 (vs. .46 actually gained) and the seniors .05 (vs. .67 actually gained). If these gains are retained until the end of their university education, these students will graduate with better critical thinking skills than they would have had without such a stand-alone

critical thinking course. Paradoxically enough, the greatest benefit from taking this course seems to have accrued to the students in Levels 4 and 5 who were in the last semester of their undergraduate education; their numbers, however, are too small to permit meaningful generalizations.

In combination, these results point to the following generalizations: One semester of university education, without a course dedicated to teaching critical thinking, will improve a student's critical thinking skills very little, especially after the first year. A traditional one-semester critical thinking course without computer-assisted instruction will improve them a little. A one-semester critical thinking course relying solely on computer-assisted instruction and machine-scored multiple-choice tests will improve them a little bit more. But a one-semester critical thinking course which combines computer-assisted instruction with graded written assignments and tests will improve them the most.[18] In no case, however, will the improvement be very great; the maximum improvement that can be expected after a one-semester course is an average increase of a little less than one standard deviation of the students' performance on a pre-test.

The results *point to* these generalizations, but of course they do not prove them. Many other explanations of the results are possible, and will be considered below. In order to test the proposed generalizations, many more studies of a similar kind need to be done, with a standard form of reporting which facilitates comparisons and makes it possible for instructors to adopt course designs which prove to be especially effective and efficient.

The following alternative explanations of the differences reported in these studies occur to me. (1) Scores on the CCTST might be a poor measure of the critical thinking skills which

critical thinking courses are designed to improve. (2) Groups may differ with respect to the incentives they had to perform well on the pre-test, or with respect to the incentives they had to perform well on the post-test. (3) Groups may differ with respect to the amount of work done in the course at the time they took the post-test. (4) Groups may differ in composition in ways that affect the degree to which they can improve their critical thinking skills in a single semester. (5) The formats and manner of instruction of the critical thinking courses may differ in other causally relevant respects than whether they had computer-assisted instruction and whether they had written graded assignments and tests. (6) The greater improvement in courses using computer-assisted instruction might have been due to improvements in general ability to answer multiple-choice items, because of the extensive practice in answering such items which those in traditional courses did not get.

These alternative explanations are of course not mutually exclusive. The differences in mean gains in the different studies may be due to a combination of such factors, or to a combination of them with the main differences in course format. Before considering each alternative possible explanation in detail, however, we should note that at least one explanation can be ruled out: the differences are not due to whether students wrote the same form of the CCTST on the post-test as on the pre-test; the present study has shown that it makes no difference to one's gain on this test whether one writes the same form or a different form.

Let us now consider the six alternative possible explanations in detail.

(1) *Invalid test?*: As the test manual (Facione et al. 1998) and first technical report (Facione 1990a) indicate, the California Critical Thinking Skills Test is based on an expert consensus statement of the critical thinking skills which might be expected of college freshmen

and sophomores (American Philosophical Association 1990) The statement, endorsed by a panel of 46 persons active in critical thinking education, research and assessment (including the present author) after a two-year-long, multi-stage Delphi process, grouped critical thinking skills[19] into six groups, as follows:

1. interpretation: categorization, decoding significance, clarifying meaning
2. analysis: examining ideas, detecting arguments, analyzing arguments into their component elements
3. evaluation: assessing claims, assessing arguments
4. inference: querying evidence, conjecturing alternatives, drawing conclusions
5. explanation: stating results, justifying procedures, presenting arguments
6. self-regulation: self-examination, self-correction

The 34 items on Form A of the CCTST were selected from a bank of 200 previously piloted multiple-choice items on the grounds of their apparent clarity, level of difficulty and discrimination. Form B was created after the studies which validated Form A by varying either the content of each question or the order of possible answers, or both, but keeping the form of each question the same. The 34 items target five of the six types of critical thinking skills: interpretation (5-9), analysis (10-13), evaluation (1-4), inference (14-24) and explanation (25-34).[20] When one looks at the items in each group, however, one finds that they target only some of the skills distinguished by the Delphi panel. Each of the five items in the interpretation group (5-9), for example, requires selection of a statement equivalent to a given statement; if this activity exemplifies “decoding significance”, then no items on the CCTST test skills of categorization or clarifying meaning. In the analysis group (10-13), one item requires recognition

that a passage contains no argument, one requires identification of the main conclusion in an argumentative passage, one requires recognition of the argumentative role in the same passage of a given statement, and one requires identification of the missing premiss in a brief argument; thus no item in this group tests the skill of "examining ideas" which the Delphi panel distinguished from the skills of detecting and analyzing arguments. In the evaluation group (1-4), all four items require determination of whether the conclusion of a simple (single-inference) argument follows from the given premisses, either necessarily or with probability, or whether on the other hand the contradictory of this conclusion follows either necessarily or with probability; thus no item in this group tests the skill of assessing claims, and in particular there is no item anywhere on the CCTST which tests the skills of assessing the credibility of observation reports or of alleged expert opinions, skills which most persons active in critical thinking research, education and assessment would take to be central critical thinking skills. In the inference group (14-24), items generally require identification of which one of a given set of statements follows, either necessarily or with probability, from given information; no items test the skill of querying evidence, and only one (item 21) tests the skill of conjecturing alternatives. In the explanation group (25-34), nine items (26-34) test in various ways the skill of identifying the fallacy in a piece of reasoning and the remaining one (25) tests the skill of figuring out how to disconfirm a hypothesis; while this latter item can be construed as testing the skill of justifying procedures, it is hard to see how the others test any of the component skills of explanation identified by the Delphi panel: stating results, justifying procedures, presenting arguments. The omission of items specifically testing the sixth group of self-regulation skills (self-examination and self-correction) is not a flaw, because these skills are required for answering all the items; self-regulation is

definitive of critical thinking generally.

It is possible, of course, that the skills not tested by any items in the CCTST correlate so well with the other skills which are tested that a person's score on the CCTST is a good measure of all their critical thinking skills. But such correlations would have to be demonstrated. It is also possible that the Delphi panel's list of skills requires additions, subtractions or modifications of items in the light of its general conception of critical thinking or some refinement of that conception. One would get a somewhat different test, for example, if one worked from the list of skills developed over several decades of research by Robert Ennis (Ennis 1985, 1991) or from the proposal of Alec Fisher and Michael Scriven to define critical thinking as the "skilled, active interpretation and evaluation of observations, communications, information and argumentation" (Fisher and Scriven 1997: 20).

There are also legitimate questions about the soundness of some of the items on the CCTST.[21]

Among the interpretation items, one (item 5) has two correct answers among the options. The item requires selection of a statement equivalent to a given statement of the form "Not all As are B". Two options have the form "All As are not B" and "Some A is not B", with the content for "A" the same throughout, and likewise the content for "B". The latter option is the keyed answer. But in most people's usage the former option means the same as "It is not the case that all As are B" rather than "All As are non-Bs"; thus in most people's usage the former option is also a correct answer. The defect in this item could be repaired by modifying "All As are not B" to "All As are non-B" (or "un-B" or "in-B" or "im-B", depending on the content for "B").

Another interpretation item (item 7) has two correct answers among the options on Form

A. This item requires selection of the best interpretation of a statement of the form "The K offers many Ms" (as in "The maple offers many shades of leaves"). The keyed answer is a corresponding statement of the form "Not every K has the same M". But the initial statement is given no context, and, if a single K could have many Ms, then there are contexts in which its best interpretation would be "There is a thing that has more than one M and it is a K", which is an "incorrect" option on the test. And indeed, on one of the two Forms of the test, Form A, a single K could (and in fact in some instances does) have many Ms. The word "offers" is also peculiar in the given statement; it may have been used instead of "has" in order to block the "incorrect" interpretation, but the effect is to make the statement sound strange. A remedy for this problem would be to use "has" or some other colloquial verb but to use values for "K" and "M" such that an individual K can have only one M, e.g. "The maple tree has many heights" or "The maple tree comes in many heights."

Another interpretation item (item 9) chooses an incorrect answer as the key instead of the correct answer. This item requires selection of a statement equivalent to a given statement of the form "It is not true that if p then q". The keyed answer, which has the form "p, yet not q", assumes a truth-functional interpretation of the indicative conditional "if" in ordinary English. This interpretation is notoriously controversial. For example, most people would count as false the statement "If 8 is divisible by 2, then 8 is divisible by 4", because it does not follow from the fact that a number is divisible by 2 that it is divisible by 4. (Similarly, even if it is raining at the time the statement is made, most people would count as false the statement, "If it is cloudy now, then it is raining now", because it does not follow from the fact that it is cloudy at a given time that it is raining at that time.) Thus most people have an interpretation of statements of the form

"If p then q" according to which their denials are not equivalent to the corresponding statement of the form "p, yet not q". "It is not true that if 8 is divisible by 2 then 8 is divisible by 4" is true, but "8 is divisible by 2, yet 8 is not divisible by 4" is false. Similarly, "It is not true that if it is cloudy now, then it is raining now" is always true, but "It is cloudy now, yet it is not raining now" is sometimes false. On the analysis of indicative conditionals in ordinary English which I favour, and which captures accurately the evaluative practices of most educated native speakers of English, the correct answer to the item is the last option, "None of the above is even roughly equivalent."

In the inference group, one item (item 19) has no correct answer. The stem describes an irreflexive, asymmetric, transitive binary relation whose extension includes in the subject term only human beings and in fact includes all human beings alive today and includes in the object term some but not all humans. The relation is given an invented name. The stem then supposes that the human species has two ultimate ancestors, and asks what "we can say for sure" on the basis of all this information about the participation of these individuals in the defined relation. The keyed answer, that all humans stand in this relation to these two ancestors, assumes that only the relation of being a human descendant satisfies the description. Of course, being a human descendant does satisfy the description, but so does the relation of being a human financial superior, if we assume that even the poorest human being alive today is a financial superior of (i.e. wealthier than) some extremely poor remote human ancestor. The reader can invent other relations which satisfy the description. Since the description does not deductively imply that the relation described is that of being a human descendant, it does not follow for sure from the given information that all human beings today stand in this relation to the supposed ultimate ancestors

of the human species. For example, if we take the relation of being a human financial superior, it might be that not every human being alive today is a human financial superior of (i.e. wealthier than) each of these supposed ultimate human ancestors of the human species. In this situation, one of the other options, which is supposedly incorrect, would be true, that someone does not stand in this relation to either of the first two human beings. This item probably needs to be replaced by a completely different item which tests ability to draw inferences about relations, for example, one which requires recognition that an irreflexive and transitive relation is asymmetric.

Another item in the inference group (item 23) does not list the correct answer among the options. The stem quotes an argument with three premises concerning an irreflexive and transitive relation which could be symbolized by "<" or ">". Using "<", and collapsing the three premises into a single sequence, the argument has the form: $a < b < c < d \therefore b < e$. The question asked is: "What information must be added to require that the conclusion be true, assuming all the premises are true?" (underlining in original) The correct answer to this question is: "Not both $a < b < c < d$ and $b \nless e$." This is the weakest possible addition which will bring it about that the conclusion must be true if the premises are true. Any other addition will deductively entail this information, and so will implicitly add the specified information. For example, the addition of "$c = e$" would make it necessary that the conclusion is true if the premises are true. But "$c = e$" entails "Not both $a < b < c < d$ and $b \nless e$", as can be proved by reductio ad absurdum. Even "$c = e$", however, is not provided as an option; the keyed answer is of the form "$c < e$". Admittedly, none of the other options provided would make the argument deductively valid if it were added as a premise. The problem here is that the question is badly worded. It should read instead: "Which of the following, if true, would bring it about that the conclusion must be true if all the

premises are true?"

Among the items in the explanation group, one item (item 32) does not include the correct answer among the options. The item asks for an evaluation of a criticism of an inference from the results of a study to an ambitious, controversial, racist and emotionally charged general causal conclusion. The criticism is that the study does not take into account the impact of another factor. This is a bad reason for finding fault with the inference, because it presupposes that the other factor does have an impact, a presupposition which in this particular case is not in fact established and is certainly not a matter of general knowledge which test-takers can be assumed to have. But this evaluation of the criticism is not included among the options. Of the options provided, two come close: that it is a good reason because this other factor must be taken into account, and that it is a bad reason because it is difficult to measure the effects of this factor. The keyed answer is the first of these. The problem with this item could be solved fairly easily by changing the wording of the criticism so as to remove the presupposition, e.g. to the criticism that the study does not take into account any possible impact of the other factor.

There are similar difficulties with another item of the same type (item 33) in the explanation group. The criticism to be evaluated is that some other set of factors differentiating the two groups in the study is correlated with the supposed effect. Whether this is a good reason for finding fault with the inference depends on whether the other factors are in fact correlated with the supposed effect. So the correct answer is that this is a good reason if there is such a correlation, and a bad reason if there is not. This answer does not occur among the options. One option is that it is a good reason because there is such a correlation, even if one controls for the inferred causal factor. Another option is that it is a bad reason because the factors mentioned in

the criticism are not causally relevant to the supposed effect. To the best of my personal knowledge, it is not established whether there is in fact a correlation between these factors and the supposed effect, still less whether they are causally relevant to it. Certainly it is not general knowledge whether these claims are correct. The keyed answer, however, is that the criticism is a bad reason because the factors it mentions are not causally relevant to the supposed effect.

Items 32 and 33 are two of four items (31 through 34) relating to a passage describing a conclusion drawn from the results of a study. While it is reasonable to use such a format for testing ability to judge when a criticism of an inference is reasonable, the content of the passage chosen (which is the same in both Form A and Form B) is so emotionally charged that it risks breaking the concentration of test-takers on the questions. The passage cites an advocate of white supremacy arguing that whites are genetically more intelligent than specified groups of non-whites on the basis of a comparison of scores of high school students on a geography test. The inflammatory content is unnecessary for the test objective; a passage with the same logical structure could be substituted whose content is not so emotionally charged.

In summary, six of the 34 items are unsound on both forms of the test, and a seventh is unsound on Form A but sound on Form B; that is, about one-fifth of the items are unsound. On both forms, four items do not include the correct answer among the options, one keys an incorrect answer instead of the correct answer, and one has two correct answers among the options. The item which is unsound only on Form A has two correct answers among the options. In addition, four of the 34 items relate to an emotionally charged passage which may break the needed concentration of some test-takers.

The technical report on the content validity and experimental validation of Form A of the

CCTST (Facione 1990a) points out the grounding of the test in the theoretical construct developed under the auspices of the American Philosophical Association (1990), and notes that this theoretical construct is compatible with the conceptualization of critical thinking promulgated by the California State University system. It also reports statistically significant gains in test scores for various groups of students after a one-semester critical thinking course, and no significant gains after a semester for various groups of students not enrolled in a critical thinking course. This method of experimental validation assumes that a semester of university education which includes a critical thinking course does more to improve critical thinking skills than a semester of university education which does not. Thus there is some circularity in using improvement on the CCTST to measure the effectiveness of a course in critical thinking in improving critical thinking skills.

A second report on Form A of the CCTST (Facione 1990b) found moderately strong correlations between CCTST scores and Scholastic Aptitude Test (SAT) verbal scores (.55 to .62) and SAT math scores (.44 to .48), and weak correlations with college grade-point average (GPA) (.20 to .29) and the number of semesters of high school preparatory English (.13 to .19). These correlations, which were statistically significant in groups ranging in size from 184 to 473 students, provide some indication that CCTST scores measure cognitive skills of some kind. But the overall evidence for the validity of the CCTST is not strong.

(2) *Differential incentives?*: Students in the present study had an equal incentive to do well on the pre-test and the post-test, since the better of the two marks counted for 5% of their final grade in the course. Since the pre-test mark would count even if the student did not write the post-test, however, there was little reason for those students who believed that they would do no

better or who were happy enough with their pre-test mark to write the post-test.[22] Such students would fall into four groups: a small number who did very well on the pre-test and could not do much better, a larger number with consistently low marks throughout, a perhaps moderate-sized group who did well enough on the pre-test in terms of the grade they were looking for in the course, and a small remainder who thought that a small improvement in their final mark was not worth the time it would take to write the post-test. The failure of such students to write the post-test might introduce a bias, although it is impossible to estimate its direction. The fact that the pre-test mean score of students who also wrote the post-test was almost exactly the same as that of students who did not (17.03 compared to 17.35) suggests that any bias which occurred because not all registered students wrote the post-test is slight.

Facione (1990a) administered the CCTST to students who had no incentive to do well on either the pre-test or the post-test; in neither case did their score on the test count towards their grade in the course in which it was administered. Facione personally administered a November post-test and a February pre-test to more than 80% of the sections in this large study of 1,169 college students. He reports that the pre-test students seemed more cooperative and appeared to put forth a stronger effort, whereas the post-test students, "pressed at the end of the semester with a variety of deadlines and knowing that the CCTST would not influence their final course grade, although willing to participate, seemed to do hastier work and put forth less effort..." (Facione 1990a: 13) Although he does not say so, one presumes that students writing the May post-test displayed similar behaviour. These observations suggest that the post-test scores in Facione's study are low and that his students improved more than his numbers show in the skills measured by the CCTST.

The students studied by Hatcher (1999, 2001) had no incentive to do well on the pre-test, except for being urged to do their best for the sake of the validity of the study; the pre-test mark did not count towards the grade in their course. The post-test, on the other hand, counted for 10% of the final examination grade, which in turn counted for 25% of the grade for the course; thus the post-test mark counted for 2.5% of the total grade in the course. The difference in incentives might mean that the pre-test scores in Hatcher's on-going study are an under-estimate, but any such bias is likely to be small, given the tendency of students, noted by Facione, to make a good effort at the beginning of term; in fact, Hatcher's subjects were entering freshmen at the very beginning of their first year of university study. Further, the incentive of Hatcher's subjects to do well on the post-test was somewhat mitigated by the fact that the mark out of 34 on the CCTST was treated as a mark out of 20, with any score of 20 or more getting 100% for this part of the final exam; thus there was no grade incentive to do better than 20 (e-mail communication, 3 September 2002).

Van Gelder's studies (Donohue et al. 2002) used the same incentives as the present study; students got 5% of their final grade for the better of the two marks on the pre-test and the post-test. Thus his results are comparable in this respect to those of the present study.

(3) *Different timing of the post-test?*: It might make a difference to the mean gains whether one administered the post-test in the last week of classes or after the students had studied for the final examination. The present study, like those by Facione (1990a) and van Gelder (Donohue et al. 2002), administered the post-test in the last week of classes, before students had begun studying for the final examination. So did the students in Hatcher's studies (e-mail communication, 16 April 2003). If studying for the final examination in a critical thinking course

improves critical thinking skills, the post-test scores in the present study and in those by Facione, van Gelder and Hatcher underestimate the students' critical thinking skills at the end of the courses they were taking. There are however no data available on whether and how much critical thinking skills improve between the last week of a critical thinking course and the writing of a final examination.

(4) *Differential composition?*: It is theoretically possible that differences in age, sex, ethnicity, extent and type of previous education, verbal aptitude, mathematical aptitude or other variables are correlated with the extent of improvement shown after critical thinking instruction. If so, differences in the distribution of such variables could explain differences between the gains shown by the students in the present study and those shown by students in other studies. Facione (1990b: 9-10) reports a strong correlation between native language and gain in CCTST scores after critical thinking instruction: whereas native speakers of English had a mean pre-test score of 16.65 and a mean post-test score of 17.59, for a gain of 0.94, non-native speakers of English had a mean pre-test score of 13.78 and a mean post-test score of 13.73, a loss of .05. Unfortunately, Facione made this comparison between different groups of students in each case; he compared the scores of 373 native speakers of English on a February 1990 pre-test to the scores of a different group of 388 native speakers of English on a November 1990 post-test, and of 89 non-native speakers on a February 1990 pre-test to a different group of 91 non-native speakers on a November 1990 post-test. Thus, the fact that one group of non-native speakers had a lower score at the end of a critical thinking course in one semester than another group had at the beginning of a critical thinking course in another semester might be due to the fact that the first group was weaker as a whole with respect to variables causally relevant to CCTST performance.[23] For

example, Facione reports that the mean-post-test score of the 462 students who wrote the CCTST in November 1989 after a critical thinking course was 16.83 with a standard deviation of 4.67, whereas the mean post-test score of 262 students who wrote the CCTST in May 1990 after a critical thinking course was 17.38 with a standard deviation of 4.58. The difference of .55 (or .12 of the standard deviation on the pre-test) is a substantial difference for such large groups, and indicates that the 91 non-native speakers writing the November post-test could well have been a weaker group of students than the 80 non-native speakers writing the February 1990 pre-test. Nevertheless, Facione's data give some indication that native speakers of English tend to improve their critical thinking skills (as measured by the CCTST) more than non-native speakers. If so, a group with a substantial percentage of non-native speakers would be expected to improve less than a group composed almost exclusively of native speakers of English. The 278 students in the present study included a substantial percentage of non-native speakers, although quantitative data are not available; several students asked the instructor for permission to use inter-language dictionaries on the tests and examination, several students spoke English with some difficulty and with a pronounced accent in conversation with the instructor, and several messages sent to the course e-mail account were in a type of broken English characteristic of a non-native speaker. Facione (1990b: 9) reports that 19% of both his November post-test group and his February pre-test group were non-native speakers of English, a substantial percentage which is likely to have reduced the main gain from what would otherwise be expected. Information is not publicly available about the percentage of non-native speakers among the students studied by van Gelder (Donohue et al. 2002) and Hatcher (1999, 2001).

Facione (1990b) provides some evidence that age and the number of college units

previously completed are irrelevant to how much students' critical thinking skills will improve during a critical thinking course. A stepwise multiple regression analysis of factors correlated with post-test score, including the pre-test score, eliminated both age and number of college units previously completed as predictors of post-test performance, given the other factors. The present study confirms this indication for non-freshman students, in that the differences in mean gain by Level (Table 4) were not statistically significant; although the students' ages are not known, they would be strongly correlated with their Level of registration. On the other hand, Facione (1990c: 3-5) did find that sex (male or female) made a difference; although men and women did more or less equally well on his February 1990 pre-test, men did better than women on the May 1990 post-test.[24] This difference disappeared, however, when a multiple regression analysis controlling for SAT-verbal and SAT-math scores was performed; in other words, if one compares men and women in his study with the same SAT-verbal and SAT-math scores, the differences in post-test CCTST performance by sex are not statistically significant. The 278 students in the present study were not classified by sex, but the class in which they were registered had, from the instructor's visual impression, slightly more women students than men. Facione's subjects (1990a, 1990b, 1990c) were divided almost equally by sex in the second semester of 1989-90 (February to May 1990), but there was a substantial majority of women (55.2%, compared to 44.8% men) in his November 1990 post-test group of 449 students.

Some of Facione's data suggest that ethnicity might make a difference to how much a person's CCTST score improves after taking a critical thinking course. Among native speakers of English, two ethnic groups ("Blacks" and "foreigners"[25]) made significantly greater gains (2.1 and 2.0 points respectively) than Whites, who gained on average only 1.3 points (Facione 1990c:

6-7). But the Blacks and foreigners in question were very small groups (n=13 and n=7 respectively), so they may well be untypical. Further, the gain scores are confounded by the fact that post-test scores were included by ethnicity for those who wrote the November post-test but no pre-test; hence it is not the same group of students whose mean pre-test score and mean post-test score are being compared. Further, when a multiple regression analysis was performed, controlling for SAT verbal score, SAT math score and college GPA, the correlation disappeared. The evidence from Facione's study thus suggests that, among native speakers of English, ethnicity makes little if any difference to a student's change in CCTST score after taking a course in critical thinking. The students in the present study were ethnically mixed, though predominantly (from their visual appearance) of European origin. Although no individual data on these students' ethnic background is available, there were substantial "visible minorities" of students of East Asian (predominantly Chinese), Southeast Asian (predominantly Vietnamese), South Asian (predominantly Indian and Pakistani) and Middle Eastern origin, with a small sprinkling of Blacks of recent African origin (e.g. from Ghana). If ethnicity makes a difference, the students in this study are sufficiently varied that any such difference is unlikely to affect the results. No information is reported on the ethnicity of students in the studies by Hatcher (1999, 2001) and van Gelder (Donohue et al. 2002).

As for differences in distribution by academic discipline, the present study found no statistically significant difference in gains by Faculty of the students' registration. Although students from two Faculties (business and humanities) showed relatively lower gains (.10 and .37 of a standard deviation) and students from one Faculty (engineering) showed relatively higher gains (.69 of a standard deviation), the numbers in each of these three cases were too small to

make meaningful generalizations possible. In the large groups by Faculty (science and social sciences students), the mean gain was virtually identical (.54 and .47 of a standard deviation). This result conflicts with Facione's finding (1990c: 8) of a statistically significant correlation between self-reported academic major and the difference between pre-test score and post-test score; those students reporting their major as falling into the cluster "mathematics, engineering, statistics, computer science" showed the largest gain (2.04 points, or .45 of a standard deviation) and those in the cluster "natural sciences, physical sciences, health professions" showed the smallest gain (.09 points, or .02 SD). There were relatively high intermediate gains among the "letters, languages, English, liberal studies, history, humanities" cluster (1.32 points, or .29 SD) and among the "social sciences, psychology, human services, teaching" cluster (1.11 points, or .24 SD), and relatively low intermediate gains among the "business, administration, management, government, military science" cluster (.63, or .14 SD) and the "performance studies, drama, art, music, physical education" cluster (.62, or .13 SD). Although Facione's groupings do not exactly correspond to the Faculties in the present study, they are close enough to indicate that the rankings of his academic groupings with respect to mean gain reported are quite different than those in the present study. Engineering is at the top in both, but natural science at the bottom in his and second of five in the present study. Business is at the bottom in the present study, but second highest in Facione's study (with the largest number of students, 39% of the total). Humanities is second lowest in the present study, but its two equivalents in Facione's study are second highest and second lowest. Social sciences ranks third in both studies. The mix of rankings makes it impossible to draw any general conclusions about correlation of academic major with gains from taking a critical thinking course. Further, Facione's reported results are

confounded by the fact that his post-test mean is for a different group of students than his pre-test mean.[26] Further, there were statistically significant correlations between the academic majors of students in Facione's study and their SAT verbal score, SAT math score and whether they were native speakers of English, differences which could easily explain the differences in mean gain in CCTST scores by academic major. These differences, however, were not always in the direction one would expect; the academic majors with the highest mean gain (those with a major in the cluster: engineering, computer science, mathematics, statistics) also had by far the smallest percentage of native speakers of English (58%, compared to anywhere from 75% to 94% in the other clusters). While Facione's students were concentrated in the business (39%), social sciences (20%) and letters (18%) clusters, students in the present study were concentrated in the social sciences (51%) and the natural sciences (26%). There is no information on the distribution by academic major of students in the studies by Hatcher (1999, 2001) and van Gelder (Donohue et al. 2002).

Facione's data (1990a, 1990b, 1990c) indicate that verbal aptitude and mathematical aptitude are correlated with the extent of improvement in critical thinking skills after taking a critical thinking course. A stepwise multiple regression analysis found that four factors accounted for 71% of the variance in the post-test scores of 401 students: pre-test score, SAT verbal score, SAT math score, college GPA (Facione 1990b: 6). The Beta weights for these four factors in the regression equation indicated that the overwhelmingly predominant factor was the pre-test score (.70), but verbal aptitude (.13) and to a lesser degree mathematical aptitude (.08) had some importance. The correlation of verbal and mathematical aptitude with extent of improvement was also revealed by the fact that differences in mean gain by sex became statistically insignificant

once verbal and mathematical aptitude were taken into account, and differences in mean gain by ethnicity also became statistically insignificant once verbal and mathematical aptitude, college CPA and native language were taken into account (Facione 1990c: 3-8; see comments above). Thus, not only are CCTST pre-test scores correlated with verbal and mathematical aptitude,[27] but so are the gains from pre-test to post-test. This indicates that one would expect higher gains from students with higher pre-test scores, since they would tend to have better verbal and mathematical aptitude. If one looks at Table 1, it is quite striking that mean gains in one-semester courses are ranked exactly the same as the mean pre-test scores; the students in Facione's study have the lowest mean gain and also the lowest mean pre-test score, the students in van Gelder's study are highest in both respects, and the students in the present study are intermediate. Hatcher's results, although they are not strictly comparable because the students had two semesters of instruction including the writing of five argumentative essays, are all the more impressive because the students started with lower CCTST scores, and thus probably with less verbal and mathematical aptitude, than students in the other studies.

Of all the differences among the groups of students in these various studies which might explain their differential gains in critical thinking skills, verbal and mathematical aptitude have the strongest claim. The raw data, however, do not permit quantitative estimations of the explanatory role of these variables through stepwise multiple regression analysis.

(5) *Other differences in format and manner of instruction?*: The critical thinking courses taken by students in these various studies probably differed in respects other than whether they had computer-assisted instruction and whether they required students to write argumentative essays. They used different textbooks, they were taught by different instructors from different

disciplines (philosophy, psychology, reading), and so on. Not much is known about how such factors are correlated with gains in critical thinking skills after taking a critical thinking course. In a group of more than 700 students taking critical thinking courses from 14 different instructors, Facione (1990b: 11-13) found statistically significant correlations of post-test scores with only two of six instructor-related variables: number of critical thinking sections taught in the last three years and number of years of college teaching.[28] But these variables ceased to be statistically significant when a multiple regression analysis controlled for student-related factors such as SAT scores and college GPA. The scatter plot of post-test means by number of sections of critical thinking taught in the past 36 months shows no clear pattern. The scatter plot of post-test means by number of years of teaching experience shows a clear but non-linear relationship; for the first 12 years of teaching, post-test mean scores in an instructor's critical thinking class were generally higher with more teaching experience (ranging from 15.4 with one year's experience to 19.0 with 11 or 12 years' experience), but after that post-test mean scores were generally *lower* with more experience (ranging from 19.0 with 12 years' experience to 16.1 with 21 years' experience). The existence of a correlation of years of teaching experience with mean gains, at least for the first 12 years of teaching, is supported qualitatively by the decline in mean gains at Baker University over the period from 1996/97 to 2001/02; Table 1 shows a decline from a mean gain of 3.35 points to one of only 2.17 (.75 SD to .48 SD). Students in these studies take critical thinking instruction in sections of 20 taught by different instructors; Hatcher reports (e-mail communication) that Baker University has relied increasingly in recent years on adjunct instructors and new professors rather than on experienced professors. The instructor of students in the present study had at the time of the study 32 years of experience as a full-time university professor; if one can extrapolate from

Facione's results, it is a miracle that the students improved at all.

Class sizes differed in these various studies, from 20 per section in Hatcher (1999, 2001), to 25 to 30 in Facione (1990a: 11), to 42 to 117 in van Gelder's study (Donohue et al. 2002), to 400 in the present study. Whatever correlation there might be between class size and mean gain in critical thinking skills, once other explanatory factors are controlled for, it is not substantial enough to have produced any clear pattern in the present group of studies. The trend, in fact, is for larger classes to show greater mean gains. The data in these studies do not support an argument that small classes are necessary for truly effective instruction in critical thinking.

(6) *Differential training in answering multiple-choice questions?*: Since the CCTST is a multiple-choice test, and there is a knack to doing well with this test format, it is possible that a multiple-choice format for tutoring and in-course testing would give students an extra advantage on the CCTST post-test. Students in the present study had extensive practice in answering multiple-choice critical thinking items: the LEMUR software has hundreds of such items; there were hundreds more on the course Web site in the form of old assignments, tests and exams with answers; and two assignments and a mid-term test were in multiple-choice format. The students studied by Facione (1990a) and Hatcher (1999, 2001) had no practice during their critical thinking course in answering multiple-choice items. The students studied by van Gelder (Donohue et al. 2002) had an intermediate amount of practice with such items. These differences could explain partly why students in the present study and in van Gelder's study improved more on the CCTST than those in Facione's study. The data do not permit a multiple regression analysis which would show whether practice in answering multiple-choice items is significantly correlated with gains in CCTST scores once other relevant factors are controlled for. It is worth

noting, however, that all university students in North America have extensive experience in writing multiple-choice items. In the United States, for example, college-bound students must write the Scholastic Aptitude Test. When the present author asked the students in the present study how many had previously answered a multiple-choice test using the university's optical scanning cards, every single student immediately put their hand up; it was clear from the instant and emphatic reaction that this was a common form of testing them in their courses. Such facts make it unlikely that the students in the present study improved their general ability to answer multiple-choice items, and thus unlikely that this factor accounts for any of the observed gains in CCTST scores.

Thus, four factors other than course format and content very possibly account for some or all of the apparent differences in mean gains in critical thinking skills in the various studies which used the CCTST. First, there are serious questions about the "construct validity" of the CCTST (Forms A and B) as an instrument for measuring critical thinking skills, both because it does not test for all components of the conceptualization of critical thinking from which it is derived and because six or seven of the 34 items (depending on which form of the test is used) do not have exactly one correct answer among the options provided; these questions in turn imply caution about taking improvements in CCTST scores as a measure of how much critical thinking skills have improved. Second, students in Facione's studies (1990a, 1990b, 1990c) appeared to put more of an effort into the pre-test than into the post-test, whereas students in Hatcher's studies (Hatcher 1999, 2001) had incentives to do well on the post-test which were lacking on the pre-test; these imbalances in incentives could partly explain the relatively low mean gain by the students studied by Facione, and the relatively high mean gains by the various cohorts of students

studied by Hatcher. Third, differences in pre-test mean among the groups probably indicate differences in mean verbal and mathematical aptitude, factors which are strongly correlated with extent of improvement after instruction in critical thinking; such differences could partly explain the fact that the mean gain after a one-semester critical thinking course exactly correlates with the mean pre-test score, with the lowest numbers on both measures in Facione's study (Facione 1990a), the intermediate numbers in the present study, and the highest numbers in van Gelder's studies (Donohue et al. 2002); see Table 1. Fourth, the number of years of teaching experience of the instructors in these various courses might account for some of the differences in mean gain. Of the six possible factors considered, only two seem unlikely as even a partial explanation for the observed differences in mean gain in these students: differences in timing of the post-test and differential training in answering multiple-choice questions.

To say that it is quite likely that the four factors collectively account for at least some of the differences in mean gains in the different groups is not to say that they actually do account for some of that difference. Further, if they do account for some of the difference, we do not know how much. To answer such questions, more research is needed. In particular, there is an urgent need to investigate further the validity of the California Critical Thinking Skills Test, for example by administering it along with some other test of critical thinking skills.[29] The Ennis-Weir Critical Thinking Essay Test (henceforth "the Ennis-Weir") might be a good candidate for such a comparison, since it closely replicates the actual practice of critical thinking: test-takers read a letter to the editor divided into paragraphs, and write a critique of the letter paragraph by paragraph, followed by a paragraph in which they summarize and justify their evaluation. For a description of the test, see Norris and Ennis (1989: 80-84). If CCTST scores prove not to

correlate well with Ennis-Weir scores, then researchers should either use the Ennis-Weir itself, which has strong construct validity, or develop an easily scored multiple-choice test with a good correlation with the Ennis-Weir. The disadvantage of using the Ennis-Weir is that it requires manual grading of the essay-type answers by carefully trained graders, with more than one person grading each answer so as to control for inter-grader reliability. At 10 minutes per test per grader, grading is expensive.

There is also a need, once a well-validated instrument for measuring critical thinking skills has been found, to replicate the present study with different groups of students taking critical thinking courses with different course content from different instructors using different methods. In such studies it would be desirable to measure and report on the following factors: the topics covered in the course, the textbook used, the types of work used to determine the students' grade in the course (in particular the balance between essay-type questions, short-answer questions, and multiple-choice items), class size, the number of years of teaching experience of the instructor(s), the Level of registration of the students, the verbal and mathematical aptitude of the students, the percentage of students whose mother tongue is not English, the instrument used at pre-test and post-test, any incentives students had to do well on the pre-test or the post-test, and the stage of the course at which the post-test was given.

5. Conclusions and summary

Non-freshman university students taking a 12-week critical thinking course in a large single-section class of 400 students, with computer-assisted guided practice as a replacement for small-

group discussion, and all testing in machine-scored multiple-choice format, improved their critical thinking skills, as measured by the California Critical Thinking Skills Test (Forms A and B), by .49 of a standard deviation. Allowing for an expected improvement of about .05 standard deviation without a critical thinking course, this improvement is a moderate effect. It is greater than the increase of .32 of a standard deviation reported for a group of traditional-format one-semester courses in critical thinking, but smaller than increases ranging from .73 to .89 of a standard deviation reported over an 11-week interval in a course combining computer-assisted instruction and essay assignments, and generally smaller than increases ranging from .46 to .75 of a standard deviation in various cohorts of a 32-week course combining critical thinking instruction with the writing of argumentative essays. The combination of results suggests that a one-semester undergraduate course in critical thinking may be least effective in improving critical thinking skills if it uses only a traditional format without computer-assisted instruction, effective to an intermediate degree if it uses only computer-assisted instruction and multiple-choice testing, and most effective if it uses both computer-assisted instruction and essay-type assignments. But this conclusion can only be entertained as a tentative explanatory hypothesis. More work needs to be done to check the validity of the CCTST as a measure of critical thinking skills. And the present design needs to be replicated in other settings in order to quantify the extent to which these observed differences in improvement in different types of courses can be explained by differences in verbal and mathematical aptitude among the different groups of students, differences in years of teaching experience of the different instructors, differences in quality practice with argument mapping, and differences in study design with respect to the incentives to do well at pre-test and post-test and the stage of the course at which the post-test

was administered.

Acknowledgements

I would like to express my thanks to Mary Lou Schmuck, who did the statistical analysis of the results reported above, and also to Geoff Norman, who acted as a consultant on the design of the study and on the statistical analysis, and who generously allowed his research assistant, Mary Lou Schmuck, to do the analysis. I also express my appreciation to Don Hatcher for lending me his copies of Form A of the CCTST and for providing extra data, to Tim van Gelder and Don Hatcher for their comments on drafts of this paper, to Geoff Cumming for answers to questions about measuring effect size and its confidence interval, and to Peter Facione for generously providing extra data. Any errors in the present paper are my responsibility.

References

American Philosophical Association (1990). Critical thinking: a statement of expert consensus for purposes of educational assessment and instruction. *ERIC* document ED-315423.

Cohen, Jacob (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2nd edition. Hillsdale, NJ: Lawrence Erlbaum.

Donohue, Angela, Tim van Gelder, Geoff Cumming, and Melanie Bissett (2002). Reason! Project Studies, 1999-2002 (Reason! Project Technical Report 2002/1). Melbourne: Department of Philosophy, University of Melbourne. Downloadable from

http://www.philosophy.unimelb.edu.au/reason/.

Ennis, Robert H. (1962). A concept of critical thinking: a proposed basis for research in the teaching and evaluation of critical thinking ability. *Harvard Educational Review* 32: 81-111.

Ennis, Robert H. (1997). A taxonomy of critical thinking dispositions and abilities. In Joan Boykoff Baron and Robert J. Steinberg (eds.), *Teaching Thinking Skills: Theory and Practice* (New York: W. H. Freeman), 9-26.

Ennis, Robert H. (1991). Critical thinking: a streamlined conception. *Teaching Philosophy* 14: 5-24.

Facione, Peter A. (1990a). The California Critical Thinking Skills Test: College Level, Technical report #1 – Experimental validation and content validity. *ERIC* document ED 327549.

Facione, Peter A. (1990b). The California Critical Thinking Skills Test: College Level, Technical report #2 – Factors predictive of CT skills. *ERIC* document ED 327550.

Facione, Peter A. (1990c). The California Critical Thinking Skills Test: College Level, Technical report #3 – Gender, ethnicity, major CT self-esteem and the CCTST. *ERIC* document ED 327584.

Facione, Peter A. (1990d). The California Critical Thinking Skills Test: College Level, Technical report #4 – Interpreting the CCTST, group norms and sub-scores. *ERIC* document ED 327566.

Facione, Peter A., Noreen C. Facione, Stephen W. Blohm, Kevin Howard and Carol Ann F. Giancarlo (1998). *Test Manual: The California Critical Thinking Skills Test*, 1998 revised edition. Millbrae, CA: California Academic Press.

Fisher, Alec, and Michael Scriven (1997). *Critical Thinking: Its Definition and Assessment*. Point Reyes, CA: Edgepress/ Norwich, UK: Centre for Research in Critical Thinking.

Hatcher, Donald (1999). Why critical thinking should be combined with written composition. *Informal Logic* 19: 171-183.

Hatcher, Donald (2001). Why Percy can't think: A response to Bailin. *Informal Logic* 21: 171-181.

Jacobs, Stanley S. (1995). Technical characteristics and some correlates of the California Critical Thinking Skills Test, Forms A and B. *Research in Higher Education* 36: 89-108.

Jacobs, Stanley S. (1999). The equivalence of Forms A and B of the California Critical Thinking Skills Test. *Measurement and Evaluation in Counseling and Development* 31: 211-222.

Norman, Geoffrey R., Jeff A. Sloan and Kathleen W. Wyrwich (2003). Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. *Medical Care* 41: 582-92.

Norris, Stephen P., and Robert H. Ennis (1989). *Evaluating Critical Thinking*. Pacific Grove, CA: Midwest Publications.

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

Pascarella, Ernest T., and Patrick Terenzini (forthcoming). *How College Affects Students Revisited: Research from the Decade of the 1990s.* San Francisco: Jossey-Bass.

Twardy, Charles R. (forthcoming). Argument maps improve critical thinking. Forthcoming in *Teaching Philosophy*.

van Gelder, T. J. (2000). Learning to reason: A Reason!Able approach. *Proceedings of the Fifth*

Australasian Cognitive Science Society Conference, Melbourne Jan 2000. Singapore: World Scientific.

van Gelder, T. J. (2001). How to improve critical thinking using educational technology. In G. Kennedy, M. Keppell, C. McNaught and T. Petrovic (eds.), *Meeting at the Crossroads: Proceedings of the 18th Annual Conference of the Australasian Society for Computers in Lerning in Tertiary Education*, pp. 539-548. Melbourne: Biomedical Multimedia Unit, The University of Melbourne.

Table 1: CCTST scores at pre-test and post-test							
Location	Year	n	Intervention	Pre-test (mean ± SD)	Post-test (mean ± SD)	Mean gain (score)	Mean gain (in SD[30])
McMaster University	2001	278	12-week CT course with LEMUR	17.03 ± 4.45	19.22 ± 4.92	2.19	0.49
California State University at Fullerton	1990	90	1-semester course in intro philosophy (control group)	15.72 ± 4.30[31]	16.35 ± 4.67	0.63	0.14
California State University at Fullerton	1990	262	1-semester courses in critical thinking (Psych 110, Phil 200 & 210, Reading 290)	15.94 ± 4.50	17.38 ± 4.58	1.44	0.32
Baker University	1996/ 97	152	2-semester course: CT with writing	15.14 ± 4.46[32]	18.49 ± 4.30	3.35	0.75
Baker University	1997/ 98	192	2-semester course: CT with writing	14.50 ± 3.84	17.17 ± 4.40	2.67	0.60
Baker University	1998/ 99	171	2-semester course: CT with writing	15.81 ± 4.60	17.90 ± 4.72	2.09	0.46
Baker University	1999/ 2000	153	2-semester course: CT with writing	15.91 ± 4.20	18.28 ± 4.30	2.37	0.53
Baker University	2000/ 01	184	2-semester course: CT with writing	16.22 ± 4.22	18.52 ± 4.23	2.30	0.51
Baker University	2001/ 02	198	2-semester course: CT with writing	15.30 ± 4.11	17.47 ± 4.44	2.17	0.48
University of Melbourne	2000	50	11-week CT course with Reason!Able	19.50 ± 4.74	23.46 ± 4.36	3.96	0.88
University of Melbourne	2001	61	11-week CT course with Reason!Able	18.11 ± 4.86	22.09 ± 4.27	3.98	0.89

Monash University	2001	61	1-semester course in intro philosophy (control group)	19.08 ± 4.13	20.39 ± 4.63	1.31	0.29
Monash University	2001	174	6 weeks philosophy + 6 weeks traditional CT	19.07 ± 4.72	20.35 ± 5.05	1.28	0.28
University of Melbourne	2001	42	1-semester philosophy of science (control group)	18.76 ± 4.04	20.26 ± 6.14	1.50	0.33
University of Melbourne	2002	117	11-week CT course with Reason!Able	18.85 ± 4.54	22.10 ± 4.66	3.25	0.73

The gains over one semester at McMaster were substantially greater than those in various control groups (Facione 1990 and e-mail communication; Donohue et al. 2002), and intermediate between those in several one-semester courses in critical thinking at Cal State Fullerton (Facione 1990a) and those in a one-semester course at the University of Melbourne which combined computer-assisted instruction with written graded assignments and tests (Donohue et al. 2002). The gains at Baker University (Hatcher 1999, 2001: 180, e-mail communication) are not strictly comparable, because they were measured over two semesters, during which one would expect full-time university students to show more improvement than the control group, independently of taking a critical thinking course. All groups studied were first-year students, except for the group in the present study (who were in their second, third and fourth years) and the group in the critical thinking courses at California State University at Fullerton (who were in first, second, third and fourth years). Since other studies (reported in Pascarella and Terenzini forthcoming) have found much greater gains in critical thinking skills in the first year than in subsequent years, independently of taking a course in critical thinking, the net effect of the course is

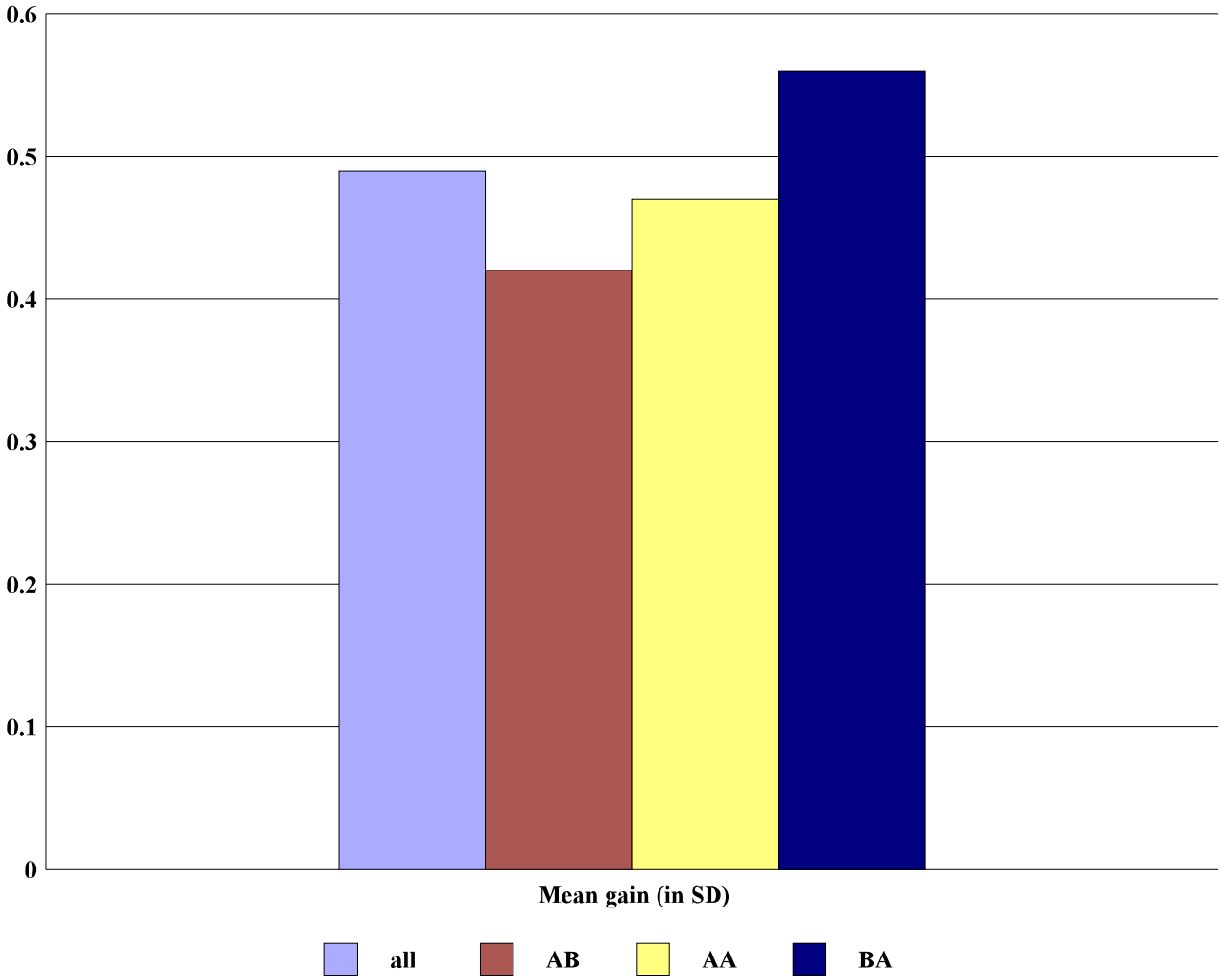
correspondingly greater for the Fullerton and McMaster critical thinking students than for the other critical thinking groups. For details of the educational interventions, consult the sources mentioned.

Table 2: CCTST scores at McMaster by form type					
Group	n	Pre-test (mean ± SD)	Post-test (mean ± SD)	Mean gain (score)	Mean gain (in SD[33])
entire pre-post	278	17.03 ± 4.45	19.22 ± 4.92	2.19	0.49
AB	90	17.34 ± 4.59	19.22 ± 4.75	1.88	0.42
AA	79	16.45 ± 4.30	18.56 ± 4.94	2.11	0.47
BA	108	17.20 ± 4.45	19.73 ± 5.04	2.53	0.56
BB	1	17	17	0	0
all preA	240	16.96 ± 4.47	n.a	n.a.	n.a.
all preB	134	17.38 ± 4.54	n.a	n.a	n.a

Gains were slightly higher among students who wrote form B first, slightly lower among those who wrote form B second, and in between among students who wrote form A twice. The differences in mean gains by form type are not statistically significant; $F(2,274) = 0.78$ ($p=.45$). But they suggest that form B is slightly harder and that it makes no difference to the post-test score whether one takes the same form as the pre-test or the other form. The mean score of all 134 students writing form B at pre-test (including 25 who did not write the post-test) was higher than the mean score of all 240 students writing form A at pre-test (including 71 who did not write the post-test), even though form B is slightly more difficult. This suggests that the students writing form B at pre-test had slightly better critical thinking skills at the outset of the course than those who wrote form A at pre-test. Further, among those who wrote the post-test, the BA group appears to have improved slightly more than the AB and AA groups, if one corrects for the extra

difficulty of form B. Allowing for the extra difficulty of form B, this group started off with slightly better critical thinking skills (as measured by the CCTST) than each of the other two groups, and it improved slightly more than they did.

Figure 1: Mean gain (in SD) by group



The AB students (second column) who wrote form A on the pre-test and form B on the post-test had a slightly smaller mean gain than the AA students (third column) who wrote form A twice, and they in turn had a slightly smaller mean gain than the BA students who wrote form B on the pre-test and form A on the post-test. The differences were not statistically significant; $F(2,274) = 0.78$ ($p=.45$).

Table 3: CCTST scores at McMaster by students' Faculty					
Faculty	n	Pre-test (mean ± SD)	Post-test (mean ± SD)	Mean gain (score)	Mean gain (in SD)
business	9	17.77 ± 4.38	18.22 ± 6.41	0.45	0.1
engineering	27	16.07 ± 3.88	19.18 ± 4.54	3.11	0.69
humanities	28	17.21 ± 5.96	18.89 ± 5.10	1.68	0.37
science	72	18.34 ± 4.35	20.75 ± 5.03	2.41	0.54
social sciences	142	16.47 ± 4.15	18.59 ± 4.69	2.12	0.47
total	278	17.03 ± 4.45	19.22 ± 4.92	2.19	0.49

There were statistically significant differences by Faculty in the students' pre-test scores; $F(4,273) = 2.79$ ($p=.02$). But the differences in mean gains by Faculty were not statistically significant; the interaction of student's Faculty by time yields an F value of $F(4,273) = 1.13$ ($p=.34$). The numbers enrolled in business, engineering and humanities are too small to make generalizations meaningful for those Faculties.

Table 4: CCTST scores at McMaster by students' Level					
Level	n	Pre-test (mean ± SD)	Post-test (mean ± SD)	Mean gain (score)	Mean gain (in SD)
1	23	14.52 ± 3.50	16.47 ± 4.03	1.95	0.43
2	158	17.05 ± 4.53	19.26 ± 4.79	2.21	0.49
3	79	17.65 ± 4.29	19.72 ± 5.29	2.07	0.46
4 and 5	17	17.41 ± 4.83	20.41 ± 4.61	3.00	0.67
total	278	17.03 ± 4.45	19.22 ± 4.92	2.19	0.49

The lower mean pre-test score of the Level 1 students reflects the fact that they were academically weak students: almost all of them (14 of 17) had completed at least a year of university studies but had not earned enough credits to advance to Level 2. In the other academic levels, mean pre-test scores tended to increase slightly as the Level increased. This pattern is consistent with findings in other studies that critical thinking skills improve slightly after the first year of undergraduate education as students take more university courses, even if they do not take a stand-alone critical thinking course. The differences in pre-test score by Level were statistically significant [$F(3,273) = 3.57$ ($p=.01$)], but were not statistically significant once the Level 1 students were excluded [$F(2,252) = 0.60$ ($p=.54$)]. The differences in mean gain by Level were not statistically significant; $F(3,273) = 0.33$ ($p=.80$). The greater mean gain by the Level 4 and 5 students may be a matter of chance, given the small number of such students. (Enrolments by students' Level add up to 277 rather than 278, because one student taking the course for credit at

another university was not classified by Level.)

Table 5: CCTST scores at McMaster by type of item					
Type of item	n	Pre-test (mean ± SD)	Post-test (mean ± SD)	Mean gain (score)	Mean gain (% of perfect score)
Definitely taught	25	11.93 ± 3.63	13.54 ± 3.92	1.61	6.44%
Possibly taught	7	3.53 ± 1.26	4.00 ± 1.47	0.47	6.71%
Definitely not taught	2	1.56 ± 0.55	1.68 ± 0.50	0.12	6.00%
All	34	17.03 ± 4.45	19.23 ± 4.93	2.2	6.47%

Mean gains by item type, expressed as a percentage of the total possible score on items of the given type, did not differ according to whether the item tested a skill judged to be definitely taught in the course, possibly taught in the course, or definitely not taught in the course. All data are for 277 students who wrote both the pre-test and the post-test, excluding the student who wrote form B at both pre-test and post-test.

Notes

1. This assignment was distributed in class, and students were allowed to work on it in groups before handing in their answers. In subsequent offerings of the course, the assignment was distributed a week beforehand, thus freeing an additional class for instruction.

2. All results are reported to two decimal places, without rounding up. Results of other researchers are reported as they stated them, except that results reported to more than two decimal places are reported to two decimal places, without rounding up.

3. The mean gain divided by the estimated standard deviation in the population is a standard measure of effect size known as Cohen's d (Cohen 1988: 20). The population from which this sample was taken can be conceptualized as: undergraduate students in English-language universities who have not yet taken a course in critical thinking. To get an estimate of the standard deviation in this population, I have used the standard deviation reported in the test manual (Facione et al. 1998: 12) for 781 undergraduate students at California State University Fullerton who took Form A of the CCTST in 1989 or 1990. This figure has the merit of being available for any study of this kind. It is close to the standard deviation at pre-test of the students in the present study (4.50 among all 374 students writing the pre-test) and also to the standard deviations at pre-test reported by Hatcher (ranges from 3.84 to 4.60) and by van Gelder (ranging from 4.04 to 4.86). It does not make sense to use a pooled standard deviation (the square root of the mean of the variances in the population at pre-test and the population at post-test), since most studies of this kind report a larger standard deviation at post-test than at pre-test. The increase in the standard deviation from pre-test to post-test reflects a spreading out of the scores, presumably as a result of higher initial scorers making larger gains than lower initial scorers. Hence the

population of undergraduate students who have taken a course in critical thinking probably has a different standard deviation in CCTST scores than the one from which the students in this sample were selected. The effect size should be measured in terms of the population who received the treatment, as they were before the treatment, not as they were after it. In this case, the effect of the critical thinking instruction is the mean gain minus the gain expected if the students had taken a full load of courses for one semester which did not include a critical thinking course. For qualitative interpretation of the effect size, see section 4, "Discussion".

4. The data were analyzed using a repeated measures analysis of variance (ANOVA) test. The within-subject repeated factor of time yields an F value of $F(1,277) = 97.18$ ($p=.00$).

5. Jacobs (1995: 94, 1999: 214) also reports higher mean scores on form A than on form B. In two successive years, he administered the CCTST to a large group of incoming freshmen randomly allocated to write forms A and B. In 1993, the mean score on part A was 16.01 ($n=684$), and on part B 15.36 ($n=692$); the following year on part A it was 15.64 ($n=753$) and on part B 15.32 ($n=708$). The differences in mean scores by form type of .65 the first time and .32 the second time would correspond to differences of 1.30 and .64 between a gain in a BA group and a gain in an AB group; thus, the difference in the present study of .65 between the BA group's gain and the AB group's gain confirms Jacobs' results. The confirmation is weak, because the present study used a smaller sample and allocation by form type was not random. (The students who wrote form B at pre-test were the students with family names beginning with the letters A through L inclusive who had heeded the instruction in the first class to go to another room for their pre-test. The students writing form A at pre-test included all those with family names beginning with the letters M through Z, along with other students who either did not hear

or did not heed the instructions; curiously, the former group did worse (mean 16.72) than the latter (mean 17.61). The students who wrote form A at pre-test were allocated to form A or B at post-test on the basis of their student number, with the more senior students writing form B and the more junior ones form A; in accordance with the general increase in critical thinking skills during an undergraduate education (Pascarella and Terenzini, forthcoming), the more senior group had a higher mean score at pre-test. The student who wrote form B both times disobeyed instructions the second time.)

6. The interaction of form type by time yields an F value of $F(2,274) = 0.78$ ($p=.45$).

7. The interaction of form type by pre-test score (among all 374 students writing the pre-test) yields an F value of $F(1,372) = 0.75$ ($p=.38$).

8. The analysis of variance between those who wrote both tests and those who wrote the pre-test only yields an F value of $F(1,372) = 0.35$ ($p=.55$).

9. A singe nursing student was grouped with the science students.

10. The interaction of Faculty by pre-test score (among 278 students writing both the pre-test and the post-test) yields an F value of $F(4,273) = 2.79$ ($p=.02$).

11. The interaction of student's Faculty by time yields an F value of $F(4,273) = 1.13$ ($p=.34$).

12. The interaction of academic level by pre-test score yields an F value of $F(3,273) = 3.57$ ($p=.01$). When the 23 Level 1 students were excluded from the analysis, the differences were not statistically significant; the F value was $F(2,251) = 0.60$ ($p=.54$). Only 277 of the 278 students who wrote both pre-test and post-test were classified by Level; the remaining student was taking the course for credit at another university, and was not classified by Level.

13. The student numbers of the 23 Level 1 students who wrote both the post-test and the pre-test indicated that 1 of them began studies at McMaster in 1997, 2 in 1998, 17 in 1999, and only 3 in

the 2000-01 academic year in which they participated in the present study. The 23 students were predominantly in engineering (10), with a sprinkling from science (6), social sciences (4) and humanities (3). The small number of students registered in Level 1, and the fact that most of them were in fact in their second year of undergraduate studies, are due to the course prerequisite of registration in Level 2 or above.

14. The interaction of student's level of registration by time yields an F value of $F(3,273) = 0.33$ ($p=.80$).

15. The within-subject repeated factor of time yields an F value for the 25 definitely taught items of $F(1,276) = 68.90$ ($p=.00$), for the 7 possibly taught items of $F(1,276) = 25.80$ ($p=.00$), and for the 2 definitely not taught items of $F(1,276) = 10.68$ ($p=.00$). (There are 276 degrees of freedom in the denominator rather than 277 because the results of the student who wrote form B at both pre-test and post-test were inadvertently excluded from this calculation. Inclusion of his results would obviously have very little effect on the statistics.)

16. Of these 262 students, 38 were registered in Psychology 110: Reasoning and Problem Solving, 31 in Philosophy 200: Logic, 116 in Philosophy 210: Argument and Reasoning, and 77 in Reading 290: Critical Reading as Critical Thinking (Facione 1990b: 7). Students in each course were taught in sections of 25 to 30 students each (Facione 1990a: 11), apparently without further division into discussion groups. Each course had been approved as a way of satisfying the requirement in the California state university system that every undergraduate student must take a course in critical thinking. Collectively, these four courses accounted for 85% of the instruction in such approved courses at California State University, Fullerton, at the time of Facione's study, the remainder being mainly in Speech Communication 235: Essentials of Argumentation and

Debate. Although Facione does not provide mean gains separately for the different courses, he reports (Facione 1990b: 7-8) statistically significant differences in the post-test mean by course (p=.03). The courses are listed above in order from lowest post-test mean to highest post-test mean: Psychology 110: Reasoning and Problem Solving (124 students, post-test mean 16.30), Philosophy 210: Logic (225 students, post-test mean 16.91), Philosophy 210: Argument and Reasoning (257 students, post-test mean 17.12), Reading 290: Critical Reading as Critical Thinking (121 students, post-test mean 17.85). Although these differences may be explainable by differences in student characteristics (pre-test score, SAT verbal and math scores, college GPA), the ranking of the courses by post-test mean is a salutary corrective to automatic prejudices about which discipline and which course content are most effective at improving critical thinking skills.

17. Facione (1990a) does not report the distribution of the 262 students in these courses by Level of registration, nor does he report the mean pre-test and post-scores with standard deviations for sub-groups sorted by Level. But the students came from courses at different levels, Psychology 110 (a freshman course), Philosophy 200 (a Level 2 course), Philosophy 210 (a Level 2 course) and Reading 290 (a Level 2 course), and the 877 students in his various studies for whom he had the relevant data had on average completed 71 semester units of undergraduate courses, i.e. slightly more than four semesters (60 units) of full-time undergraduate education. Hence it is likely that a substantial number of the 262 students were not freshmen. The above estimates of effect size are only approximations, because they make the simplifying and probably false assumption that the mean gain in standard deviations among the freshmen in the group of 262 students was the same as that among the others.

18. Van Gelder (2000, 2001) and Twardy (forthcoming) attribute the higher gains with van

Gelder's Reason!Able software to extensive guided practice in argument mapping provided by this software. A good test of this competing explanation would be to measure mean gains among students in a course with guided computer-assisted practice and written assignments, but without argument mapping. If van Gelder's and Twardy's explanation is correct, these main gains should be substantially lower than those observed in students using the Reason!Able software. If my explanation above is correct, these main gains should be about the same as those observed in students using the Reason!Able software.

19. The panel also endorsed a general conception of critical thinking and a list of dispositions of the ideal critical thinker. The general conception has at its core "purposeful, self-regulatory judgment", a conception similar to (though vaguer than) the influential conception proposed by Robert Ennis, "reasonable reflective thinking that is focused on deciding what to believe or do" (Ennis 1985). Detailed lists of component skills and dispositions, with criteria and standards for their possession, are of course much more useful for educational assessment and instruction than a general definition. In this respect, Ennis' work, going back to his landmark 1962 paper, is a model.

20. Facione (1990a) has a different classification, which appears to reflect a different ordering of the items on an earlier version of the CCTST. The test manual (Facione et al. 1998a) combines interpretation and analysis items into a broader category called "analysis", and evaluation and explanation into a broader category called "evaluation". The accompanying score sheet classifies items 5 through 13 under the broader category of analysis, items 1 through 4 and 25 through 34 under the broader category of evaluation, and items 14 through 24 under inference. The above classification into the narrower groups, based on inspection of the individual items, is consistent

with both the test manual and the scoring sheet.

21. For each of the items whose soundness is questioned, the criticisms apply equally to both Form A and Form B of the CCTST, unless there is an explicit indication to the contrary. Items on the two Forms are exactly parallel, in some cases differing in content but not in form, and generally having a different order of possible answers.

22. Little reason, but some. A number of students expressed interest in finding out by comparing their own pre-test and post-test scores how much they had personally improved their critical thinking skills while taking the course. Such students would have a reason for writing the post-test even if they did not expect their mark to improve very much.

23. Using techniques of multiple regression analysis, Facione (1990b: 7) determined that, if pre-test scores were ignored, the strongest predictors of post-test performance after a critical thinking course were in order SAT (Scholastic Aptitude Test) verbal score, SAT math score, college GPA (grade-point average) and high school GPA. These are of course predictors and not necessarily causes, but it makes sense to assume common causal factors determining these scores and scores on the CCTST after a critical thinking course.

24. Unfortunately, the comparison is not with exactly the same subjects. The mean score on the February 1990 pre-test of 479 students enrolled in critical thinking courses was 16.28 with a standard deviation of 5.08 for 237 men, and 15.90 with a standard deviation of 4.20 for 242 women. The difference is not statistically significant ($p=.36$). The data on the May 1990 post-test of students enrolled in critical thinking courses is reported for only 262 of these 479 students. Among them, the 128 men had a mean score of 18.00 (standard deviation not reported) and the 134 women a mean score of 16.79 (standard deviation not reported); this difference was

statistically significant (p=.01). The discrepancy could be due to differences in factors causally relevant to CCTST performance between the 134 women whose mean post-test score was reported and the 108 other women included in the pre-test mean who did not write the post-test or whose post-test scores were not included in the post-test mean. In fact, the 262 students (men and women) whose May 1990 post-test mean is reported had a pre-test mean of only 15.94, somewhat lower than the pre-test mean of 16.27 of the additional 217 students (men and women) whose scores were included only from February 1990. It is quite possible that it was the women among the 262 students who were responsible for the lower mean pre-test score of this group; if so, their gains would have been similar to those of the men among the 262.

25. Classification was based on the students' self-identification of their ethnicity on their application for admission to the California State University system. Applicants chose among 16 alternatives, which Facione grouped into American Indian (alternative 1), Asian (alternatives 7 through 11), Black (alternative 2), Hispanic (alternatives 3 through 6), White (alternatives 12 through 14, including "Pacific Islander" and "Filipino" as well as "White/Non-Hispanic"), and Foreign (alternative 15). Alternative 16, excluded from Facione's classification, was "Declines to state". Some students in his study left this question blank on their application, not even checking "Declines to state".

26. He reports the mean pre-test score for 473 students who wrote the February 1990 pre-test and the mean post-test score for 725 students who wrote the post-test in November 1989 or May 1990 (Facione 1990c: 8). Of these 725 students, 323 wrote the May 1990 post-test (1990a: 17). Thus the mean pre-test score includes the scores of 150 students whose post-test scores are not included in the mean post-test score, and the mean post-test score includes the scores of 402

students whose pre-test scores are not included in the mean pre-test score. Although the numbers are large and there are perhaps only weak systematic biases in the selection of the samples, differences between pre-test mean and post-test mean could well be due to individual differences among the different groups of students used to calculate the two means.

27. Facione (1990b: 4) reports a correlation of .55 (p=.00) between SAT verbal score and pre-test CCTST score, and a correlation of .44 (p=.00) between SAT math score and pre-test CCTST score. In each case the correlation is measured among 333 students.

28. The other four variables examined, for each of which the correlation with post-test mean score turned out not to be statistically significant, were full-time vs. part-time (p=.87), male vs. female (p=.74), tenure vs. non-tenure (p=.21) and PhD vs. non-PhD (p=.10).

29. California Academic Press, publisher of the CCTST, has replaced Forms A and B with a version called CCTST - 2000, which should therefore be the focus of any such investigation.

30. As an estimate of the standard deviation in the population, I have used the standard deviation of 4.45 reported in the CCTST manual (Facione et al 1998: 12) for 781 undergraduate students at California State University Fullerton who had not taken a critical thinking course when they took the CCTST in 1989 and 1990.

31. Peter Facione (e-mail address <pfacion@luc.edu>) provided previously unpublished standard deviations for this control group in personal e-mail communication.

32. Means and standard deviations for the Baker University groups have been recalculated from the raw data supplied by Donald Hatcher (e-mail address <donald.hatcher@bakeru.edu>). They therefore differ slightly from those reported in (Hatcher 1999) and (Hatcher 2001), and include data for additional years.

33. Assuming that one SD = 4.45 (Facione et al. 1998: 12).